TEXT FLY WITHIN THE BOOK ONLY

UNIVERSAL
LIBRARY

OU_160484

UNIVERSAL
LIBRARY

PROPERTIES OF SOLIDS

PRENTICE-HALL SERIES IN ENGINEERING
OF THE PHYSICAL SCIENCES

K. R. Wadleigh and J. B. Reswick, Editors

KOERBER *Properties of Solids*
LONG *Mechanics of Solids and Fluids*
ROHSENOW AND CHOI *Heat, Mass, and Momentum Transfer*

George G. Koerber

Electrical Engineering Department
Iowa State University

PROPERTIES OF SOLIDS

Prentice-Hall, Inc.

Englewood Cliffs
New Jersey *1962*

© 1962 by PRENTICE-
HALL, Inc., Englewood
Cliffs, New Jersey

All rights reserved. No part of this book may be reproduced in any form, by mimeograph or any other means, without permission in writing from the publisher. Printed in the United States of America. Library of Congress Catalog Card Number 62-11985
59959–C

PREFACE

The science of materials has evolved from the activities of two radically different groups; the engineers and the solid state physicists. Each of these groups was motivated by different objectives and adopted procedures tailored to these objectives. The engineer was primarily concerned with the selection and fabrication of materials to meet certain design objectives; the physicist was primarily concerned with understanding the basic physical processes that governed the behavior of the materials under different environmental conditions.

Because of the differences in their purposes, each group necessarily defined and evaluated materials properties in a manner that seemed most practical for its own purposes. Because neither was very much concerned with the activities of the other, a situation eventually developed in which there was much confusion about the precise meaning of terms and it became difficult for members of either group to use data from the investigations of the other.

In this decade, the more scientifically minded engineer has found his work impinging more and more closely upon that of the solid state physicist, and the need for a consensus about methods and terminology has become acute. This consensus is to be found in the discipline that we now call materials science. This field consists necessarily of a synthesis of the activities of both engineers and physicists, but by and large it is based upon the procedures, terminology, and philosophy of the physicist. In its essence, materials science is concerned both with predicting the response of a material to a specified set of environmental conditions, and with explaining its behavior, in the hope that if the causes of its behavior are understood, it may be possible to improve it. Both of these objectives must be based upon the

experimental determination of certain observable parameters, and it is here that the topic must be unified to a sufficient degree that it can be properly termed a science. It is not only necessary, for example, to understand the difference between an elastic constant and a technical elastic modulus, or an isothermal elastic compliance and an isentropic elastic rigidity, but it is necessary to be able to prescribe the conditions under which each is to be evaluated, and to know what auxiliary data are required to enable one of these to be computed from the other.

This is the basic purpose of this text; to bring a degree of coherence and unity into the variety of ways in which materials properties can be characterized, and to explain the ways in which materials respond to their environments in terms of the basic physical processes that govern this response. In order to achieve this purpose, it has been necessary to employ certain mathematical techniques that may have been absent from the normal background of many engineers. For this reason, the first three chapters are devoted to the development of some elementary vector analysis, matrix algebra, and tensor operations. Those who are already familiar with this material could still profit from a rather quick reading of these chapters in order to familiarize themselves with the notation used later in the text. Others may find that studying selected topics in these chapters will suffice to allow them to approach the following chapters with a degree of comfort.

It has not been the author's purpose to provide comprehensive information about each material property. Information of this type can be found in current technical journals and handbooks. Instead, it is hoped that this text will enable the reader to evaluate the data that he finds in the literature and will, perhaps, guide him in knowing when to be skeptical.

I would like to express my appreciation here to my wife, Barbara. She not only prepared the manuscript, but at the same time waged a patient battle with my tendency to use awkward, Germanic sentences. The comments and criticisms of Professor Harry J. Weiss of the Department of Mathematics at Iowa State University and of Professor A. J. Swalin of the Department of Metallurgy at the University of Minnesota have helped to make this a better book.

No textbook can be written without reference to the technical and scientific phases of our cultural heritage. Oddly enough, much of this text is a synthesis of ideas that were conceived in the period from 1870 to 1910. In this connection the names of such giants of continuum mechanics as Love, Neumann, and Voigt deserve mention.

G. G. K.

CONTENTS

CHAPTER ONE

Introduction to Vector Analysis **1**

1.1 Summation convention, 1; 1.2 Linear independence and Euclidean space, 4; 1.3 Vectors and vector operations, 6; 1.4 Physical significance of vectors, 9; 1.5 Unit vectors, 10; 1.6 Scalar and vector fields, 11; 1.7 The gradient, 12; 1.8 Line, surface, and volume integrals, 15; 1.9 The vector operator and the divergence of a vector, 16; 1.10 Applications, 17; 1.11 Summary, 21.

CHAPTER TWO

Matrix Algebra **24**

2.1 Basic definitions and operations, 24; 2.2 The unit matrix, the inverse of a matrix, and the transpose matrix, 27; 2.3 Matrices as rotation operators, 29; 2.4 The diagonalization of a matrix, 31.

CHAPTER THREE

Introduction to Tensor Analysis **38**

3.1 Introduction, 38; 3.2 Transformation properties of tensors, 40; 3.3 Tensor symmetry and tensor addition, 43; 3.4 The contraction operation, 45; 3.5 The dyadic product, 47; 3.6 Differentiation, 47; 3.7 Unit tensors, 48; 3.8 Symmetry of tensors of rank higher than two, 49; 3.9 Summary, 51.

CHAPTER FOUR

Structure of Solids **54**

4.1 Classification of solids, 54; 4.2 Crystal structure, 55; 4.3 Crystal defects, 63; 4.4 Polycrystalline materials, 69; 4.5 The structure of amorphous solids, 74.

CHAPTER FIVE

Bonding in Solids **80**

5.1 Introduction, 80; 5.2 Wave mechanics, 82; 5.3 Atomic spectra and atomic structure, 84; 5.4 The covalent bond, 94; 5.5 Resonance and exchange energy, 95; 5.6 Hybridization of bonds, 96; 5.7 Elementary band theory—bonding in metallic and covalent crystals, 97; 5.8 Zone theory, 100; 5.9 Applications of zone theory to metals, 108; 5.10 Correlation between zone theory and band theory, 109; 5.11 Ionic-covalent bonding and van der Waals bonding, 111; 5.12 Ionic bonding, 111.

CHAPTER SIX

Systematic Correlation of Properties **115**

6.1 System parameters, 116; 6.2 Class I: equilibrium properties, 116; 6.3 Mathematical nature of compliances and rigidities, 121; 6.4 Class II: properties of steady-state systems, 128; 6.5 Class III properties: systems with hysteresis, 129; 6.6 Class IV: physically irreversible systems, 132.

CHAPTER SEVEN

Structural Symmetry and Neumann's Principle **136**

7.1 Crystal classes and symmetry, 136; 7.2 Neumann's principle, 141; 7.3 Properties of materials represented by tensors of rank one, 141; 7.4 Properties that may be represented by tensors of rank two, 142; 7.5 Convention for choosing the orientation axes, 147; 7.6 Rank three tensor properties, 148; 7.7 Rank four tensor properties, 152.

CHAPTER EIGHT

Elasticity and Plasticity **155**

8.1 Strain, 155; 8.2 Elasticity, 157; 8.3 The elastic constants, 161; 8.4 Technical elastic moduli, 163; 8.5 Dynamic strain and the speed of sound, 167; 8.6 Plasticity, 172; 8.7 Modes of irreversible deformation; slip and twinning, 172; 8.8 Slip systems in single crystals and the critical shear stress, 174; 8.9 Critical shear stress of perfect single crystals, 178; 8.10 Dislocation theory and slip, 180; 8.11 Strain hardening, 181; 8.12 Alloying, 182; 8.13 Polycrystalline materials, 182; 8.14 Creep, 184.

CHAPTER NINE

Thermal Properties of Solids **187**

9.1 The concept of the phonon and the natural vibrations of a solid body, 187; 9.2 The energy of a phonon and the internal energy of a solid, 189; 9.3 Thermal expansion, 194; 9.4 Thermal conductivity, 195; 9.5 Thermal conductivity of metals, 199; 9.6 Rate processes and activation energy, 201; 9.7 Metastability, 203.

CHAPTER TEN

Electronic Properties of Solids **206**

10.1 The density of states function and the effective mass of charge carriers, 206; 10.2 The Fermi-Dirac distribution, 210; 10.3 Electrical conductivity, 213; 10.4 The mobility of charge carriers, 214; 10.5 Metals, 215; 10.6 Semiconductors, 217; 10.7 Applications, 222.

CHAPTER ELEVEN

Cross Conductivities **225**

11.1 Conductivities and resistivities, 225; 11.2 Thermoelectricity, 229; 11.3 The mechanism of thermoelectric phenomena, 234; 11.4 Applications of the thermoelectric effect, 238; 11.5 The Hall effect, 239; 11.6 Anisotropy of the Hall coefficient and magnetoresistance, 242; 11.7 Thermomagnetic and thermogalvnic effects, 245.

CHAPTER TWELVE

Dielectric and Magnetic Properties **248**

12.1 Polarization in a dielectric, 249; 12.2 Dielectric rigidities and compliances, 250; 12.3 The relaxation time, 252; 12.4 Spontaneous polarization in a crystal, 254; 12.5 Pyroelectric materials, 257; 12.6 Ferroelectric materials, 257; 12.7 Antiferroelectricity, 260; 12.8 Piezoelectricity, 260; 12.9 Diamagnetism and paramagnetism, 263; Ferro- and antiferromagnetism, 264.

Appendix A *Independent Components of Rank Three Tensor Properties Where $T_{ijk} = T_{ikj}$, 271.*

Appendix B *Semimatrices of the Independent Elastic Constants for the Various Crystal Classes, 275.*

Appendix C *Periodic Table of the Elements, 278.*

Index, *281*

PROPERTIES OF SOLIDS

I

INTRODUCTION TO VECTOR ANALYSIS

Very frequently we shall have occasion to concern ourselves with large numbers of variables, coefficients of variables, and sets of equations relating different variables. In order to treat these with economy and brevity, a special notation has been developed which uses the same symbol for a set of like variables and distinguishes between these with a subscript. For example, the familiar x, y, and z labels normally assigned to a set of Cartesian coordinates axes are replaced in this notational system by x_1, x_2, and x_3. If we wish to make a statement about a pair of these which would be true of *any* pair, then we can designate the pair as x_i, x_j, rather than designating all of the possible combinations, x_1, x_3; x_1, x_2; x_2, x_3.

To remove ambiguity, the range of the subscripts should ordinarily be indicated also. If a statement about the pair x_i, x_j is true regardless of how many variables are under discussion, the range may be omitted. However, if we are speaking of a finite set of n variables, the subscripts i and j may have any value from 1 to n, and the range would be expressed by a statement such as

$$i, j = 1 \ldots n$$

1.1 Summation convention

When we are dealing with combinations of variables with each other or with constants, the appearance of the same subscript twice in a given term requires a summation through the range of the repeated subscript. Thus

$$x_i x_i = x_1^2 + x_2^2 + x_3^2 \qquad (i = 1, 2, 3) \tag{1.1}$$

On the other hand, the product $x_i x_j$ $(i, j = 1, 2, 3)$ means any or all of the nine terms

$$\begin{array}{lll} x_1^2 & x_1x_2 & x_1x_3 \\ x_2^2 & x_2x_1 & x_2x_3 \\ x_3^2 & x_3x_1 & x_3x_2 \end{array}$$

Similarly

$$\frac{\partial x_i}{\partial y_i} = \frac{\partial x_1}{\partial y_1} + \frac{\partial x_2}{\partial y_2} + \ldots + \frac{\partial x_n}{\partial y_n} \tag{1.2}$$

Further illustrations of this convention are provided by identities

$$c_i b_i \ (i = 1, 2, 3) = c_1b_1 + c_2b_2 + c_3b_3$$

but

$$c_i b_j \ (i, j = 1, 2, 3)$$

could be any or all of the set

$$\begin{array}{lll} c_1b_1 & c_1b_2 & c_1b_3 \\ c_2b_1 & c_2b_2 & c_2b_3 \\ c_3b_1 & c_3b_2 & c_3b_3 \end{array}$$

Sometimes double subscripts are convenient on a single symbol. In the case of the above set $c_i b_j$, we might define

$$A_{ij} = c_i b_j \qquad (i, j = 1, 2, 3)$$

In such a case, $A_{ij}x_j$ means $A_{i1}x_1 + A_{i2}x_2 + A_{i3}x_3$.

There may also be cases in which more than one subscript is repeated. This might be illustrated by the case

$$A_{ij}d_{ij} \qquad (i, j = 1, 2, 3)$$

This case requires a double summation over the range of i and j. In expanded form this becomes

$$A_{ij}d_{ij} = A_{11}d_{11} + A_{12}d_{12} + A_{13}d_{13} + A_{21}d_{21} + A_{22}d_{22} + A_{23}d_{23} + A_{31}d_{31} + A_{32}d_{32} + A_{33}d_{33}$$

Now let us see how this notational system may be used to reduce the labor of handling sets of variables and/or equations. The set of equations

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ y_2 &= a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ y_3 &= a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \end{aligned}$$

for example, can be represented very concisely as

$$y_i = a_{ij}x_j \qquad (i, j = 1, 2, 3)$$

As a further illustration, the equation

$$\begin{aligned} a_{ij}a_{jk} &= 0, \qquad i \neq k \qquad (i, j, k = 1, 2, 3) \\ &= 1, \qquad i = k \end{aligned}$$

represents the nine equations

$$\begin{aligned} a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} &= 0, \qquad a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} = 0 \\ a_{11}b_{13} + a_{12}b_{23} + a_{13}b_{33} &= 0, \qquad a_{31}b_{11} + a_{32}b_{21} + a_{33}b_{31} = 0 \\ a_{21}b_{13} + a_{22}b_{23} + a_{23}b_{33} &= 0, \qquad a_{31}b_{12} + a_{32}b_{22} + a_{33}b_{32} = 0 \end{aligned}$$

$$\begin{aligned} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} &= 1 \\ a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} &= 1 \\ a_{31}b_{13} + a_{32}b_{23} + a_{33}b_{33} &= 1 \end{aligned}$$

The equation

$$T'_{ij} = a_{ik}a_{jl}T_{kl} \qquad (i, j, k, l = 1, 2, 3)$$

actually represents nine equations of which we cite the following two as examples:

$$\begin{aligned} T'_{11} &= a_{11}^2T_{11} + a_{12}^2T_{22} + a_{13}^2T_{33} + a_{11}a_{12}(T_{12} + T_{21}) \\ &\quad + a_{11}a_{13}(T_{13} + T_{31}) + a_{12}a_{13}(T_{23} + T_{32}) \\ T'_{12} &= a_{11}a_{21}T_{11} + a_{11}a_{22}T_{12} + a_{11}a_{23}T_{13} + a_{12}a_{21}T_{21} + a_{12}a_{22}T_{22} \\ &\quad + a_{12}a_{23}T_{23} + a_{13}a_{21}T_{31} + a_{13}a_{22}T_{32} + a_{13}a_{23}T_{33} \end{aligned}$$

The primary objective of this notational system is to enable us to make broad general statements that are true of all members of a set of variables or set of equations. Many of the basic principles which we shall have to enunciate can be stated concisely and unambiguously in this form. However, when we must get specific and compute numerical results, there are no shortcuts. Thus an expression such as

$$\begin{aligned} y_i &= a_{ij}x_j \qquad (i, j = 1, 2, 3) \\ a_{ij} &= a_{ji} \end{aligned}$$

might possibly represent a precise statement of some physical law, but if we hope to compute a value of y_i, given values of x_1, x_2, and x_3 and each of the factors a_{ij}, the equations must be expanded.

Shortly we shall define certain types of vector and tensor operations in terms of this notational system. From these operations we may derive certain relationships between material properties which would be almost impossible to express in extended form. For example, when we take up elasticity, we shall have to deal with transformations of the form

$$T'_{ijkl} = a_{im}a_{jn}a_{ko}a_{lp}T_{mnop}$$

Since this expression represents 81 equations, each of which has 81 terms on the right-hand side of the equation, it is obvious that we cannot regard this notational system as a mere convenience but rather as a physical necessity.

Under some circumstances it is desirable to indicate that two or more subscripts are always identical, but no summation is to be made. In these cases, the subscripts are set off with parentheses. Thus the term a_{ii} $(i = 1, 2, 3)$ indicates a summation

$$a_{ii} = a_{11} + a_{22} + a_{33}$$

whereas the term $a_{i(i)}$ $(i = 1, 2, 3)$ indicates any or all of the terms, a_{11}, a_{22}, a_{33}.

1.2 Linear independence and Euclidean space

Let the set of variables x_i be such that there are no nonzero constants α_i $(i = 1, 2, \ldots, n)$ which satisfy the equation

$$\alpha_i x_i = 0$$

Under these circumstances, the variables x_i are said to be linearly independent.

If each member of the set x_i is assigned an arbitrary value X_i $(i = 1, 2, \ldots, n)$, these n quantities are called a *point*. If a different group of arbitrary values are chosen to form another point Y_j $(j = 1, 2, \ldots, n)$, then the distance S between X_i and Y_j is defined by the equation

$$S^2 = (X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \ldots + (X_n - Y_n)^2$$

The totality of all possible points X_i, Y_j, $\ldots$, N_n forms an n-dimensional Euclidean space.

In this space it will be possible to have n mutually perpendicular lines intersecting at a point. A positive direction may be designated for each of these lines and a distance along any one line may be used to represent the magnitude of one of the variables x_i. If this space has three or less dimensions, we are able to visualize it, and in this case the aforementioned set of mutually perpendicular lines becomes the familiar set of Cartesian axes such as that shown in Fig. 1.1.

The reader may wonder why we choose this rather general way of setting up a Cartesian coordinate system. There are a number of reasons: first of all, we want to free ourselves from the limitations of a three-dimensional space; the second reason is that we may wish to work with problems whose dimensions may be quantities other than distances. It is hoped that by defining space in this general way we may make concepts such as phase space, momentum space, spaces whose dimensions are reciprocal times and

the like, seem less strange when we encounter them later. For example, we visualize the space of Fig. 1.1 by drawing an analogy between distance and the magnitude of each x_i. Actually x_i may represent a set of distances, but it could also represent a set of reciprocal distances, reciprocal times, linearly independent momenta or temperature gradients, or whatever set of quantities we may wish to work with in solving a given problem.

In addition, there is a third rather pragmatic reason. Many persons feel that they do not understand a concept until they visualize it. This leads us to draw geometric analogues for physical problems and to base derivations upon these analogues. These derivations are usually much more complex than equivalent derivations based upon algebraic operations. By defining space in general terms, we may free ourselves from geometric considerations and we will be able to arrive at important conclusions more readily through the algebraic equivalent. It is only in deference to those who must visualize that we invoke the concept of space at all. In general, we shall base all of our derivations upon algebraic identities and operations, and then will provide a subsequent geometric interpretation. This is the reverse of the usual procedure of giving a geometric interpretation to a physical problem and then deriving physical relationships from geometric considerations.

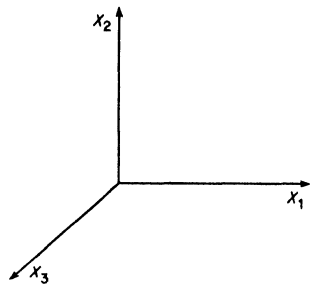


Figure 1.1 A set of Cartesian coordinates.

Given a set of linearly independent variables x_i, we may derive another set y_i by means of the linear transformation

$$y_i = a_{ij}x_j \qquad (i, j = 1, 2, \ldots, n)$$

We shall not prove this statement, but it can be shown[1] that the set y_i will be linearly independent if and only if we restrict the constants a_{ij} to values such that

$$\begin{aligned} a_{ij}a_{ik} &= q, \qquad j = k \\ &= 0, \qquad j \neq k \end{aligned}$$

If the constant q is equal to ± 1, the transformation is said to be *orthogonal*.

It is common to use a symbol called the *Kronecker delta* to specify the orthogonality condition. If we define

$$\begin{aligned} \delta_{ij} &= 1, \qquad i = j \qquad (i, j = 1, 2, 3) \\ &= 0, \qquad i \neq j \end{aligned}$$

[1] Maxime Bôcher, *Introduction to Higher Algebra* (New York: Macmillan, 1935).

then the orthogonality condition is expressed simply as

$$a_{ij}a_{jk} = \delta_{ik}$$

Now, because of the relationship between x_i and y_i, the set y_i provides an alternate characterization of the same space defined by the set x_i. If there are only three members of each set, we may visualize each in terms of a set of Cartesian coordinate axes having the same origin, but with the axes not coincident with one another. Thus the transformation defines a *rotation of axes.* (See Fig. 1.2.) The constants a_{ij} are geometrically equivalent to cosines of the angles between axes; the cosine of the angle between the y_i axis and the x_j axis being equal to a_{ij}. Let us locate a point anywhere in this space. The distance and direction from the origin O to P may be described equally well in terms of either a set X_i or a set Y_i. The magnitude of this distance is

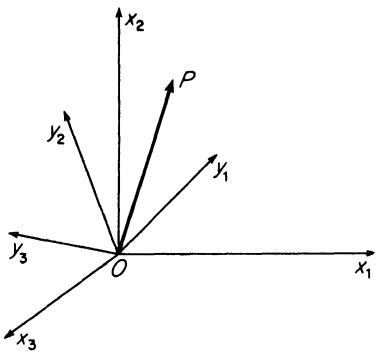


Figure 1.2 Two sets of Cartesian coordinates with coincident origin. The magnitude and direction of OP can be expressed in terms of its components along either set of coordinates, and is completely independent of the set chosen to describe it.

$$S^2 = X_1^2 + X_2^2 + X_3^2 = Y_1^2 + Y_2^2 + Y_3^2$$

and the direction is specified in terms of angles whose cosines are

$$\frac{X_i}{S} \quad \text{or alternatively} \quad \frac{Y_i}{S}$$

In other words, S, the distance and direction of the line from O to P, is an *invariant* of the space. This invariant may be described by either the set X_i or the set Y_i, and these sets are transformed into one another by the equation

$$Y_i = a_{ij}X_j \tag{1.3}$$

where a_{ij} is the cosine of the angle between the Y_i coordinate axis and the X_j coordinate axis.

1.3 Vectors and vector operations

We have shown that the location of any two points relative to one another can be specified by a distance and a direction relative to a reference frame which has meaning in the space in which we are working. We have also shown that the distance is invariant under a rotation of coordinates and that once the direction has been specified in terms of any one reference frame, it can be specified in terms of all reference frames. All of this information can be embodied in either of the ordered sets X_i or Y_i.

For our purposes, we may define a *vector* as being the ordered set X_i. This set has the property of representing a magnitude

$$S^2 = X_1^2 + X_2^2 + \ldots + X_n^2$$

a set of direction cosines X_i/S relative to a reference frame, and can be transformed through Eq. (1.3) to another equivalent set Y_i.

In general, we shall represent a vector by boldface type and use the same symbol in italic to represent its magnitude. A set of components referred to a reference frame will be represented by the same symbol with subscripts. Thus **A** is a vector whose magnitude is A and whose components referred to a predetermined set of coordinates are a_i.

Vector addition. Let **A**, **S**, a_i, and s_i be two vectors and their respective components in some coordinate system and let **B** be a vector whose components in the same coordinate system are b_1, b_2, and b_3. Then the operation defined by the identity

$$\mathbf{B} = \mathbf{A} + \mathbf{S}$$

means that $$b_i = a_i + s_i$$

This operation is commutative and associative. Thus

$$\mathbf{A} + \mathbf{S} = \mathbf{S} + \mathbf{A}$$

$$\mathbf{A} + \mathbf{S} + \mathbf{C} = \mathbf{A} + (\mathbf{S} + \mathbf{C}) = (\mathbf{A} + \mathbf{S}) + \mathbf{C}$$

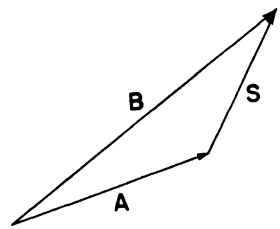


Figure 1.3 Geometric representation of the vector sum **B** = **A** + **S**.

The geometric equivalent of vector addition is illustrated for **A** + **S** = **B** in Fig. 1.3.

Vector multiplication. There are actually four distinct operations in vector algebra corresponding to multiplication. Three of these are discussed in the next paragraph; the operation known as the *dyadic product* will be taken up in Chap. 3. Let s_1, s_2, and s_3 be the components of the vector **S** in some coordinate system, and let a_1, a_2, and a_3 be components of a vector **A** in the same coordinate system. Let b be any scalar. Then the operation defined by the identity

$$\mathbf{A} = b\mathbf{S}$$

means that $$a_i = bs_i$$

In brief, **A** is a vector having the same direction as **S** but whose components are b times as large as the corresponding components of **S**.

Inner product. Let $\mathbf{A}$, a_i, $\mathbf{S}$, and s_i have the same meaning as above. Then the operation symbolized by $\mathbf{A} \cdot \mathbf{S}$ is defined by the identity

$$\mathbf{A} \cdot \mathbf{S} = a_i s_i \tag{1.4}$$

(Remember the summation convention!) This operation is sometimes called a *dot product* as well as a *scalar poduct*. $\mathbf{A} \cdot \mathbf{S}$ produces a scalar whose magnitude is $AS \cos \theta$ where θ is the angle between $\mathbf{A}$ and $\mathbf{S}$. This operation is commutative and distributive. Thus

$$\mathbf{A} \cdot \mathbf{S} = \mathbf{S} \cdot \mathbf{A}$$

$$\mathbf{A} \cdot (\mathbf{S} + \mathbf{B}) = \mathbf{A} \cdot \mathbf{S} + \mathbf{A} \cdot \mathbf{B}$$

Two vectors are said to be *perpendicular if* and *only if* their scalar product is zero.

The vector cross product and the Levi-Civita density. Let $\mathbf{A}$, a_i, $\mathbf{S}$, and s_i have the same meaning as before. Then the operation $\mathbf{A} \times \mathbf{S}$ is defined by the identity

$$\mathbf{A} \times \mathbf{S} = \mathbf{B}$$

where

$$b_1 = a_2 s_3 - a_3 s_2$$

$$b_2 = a_3 s_1 - a_1 s_3$$

$$b_3 = a_1 s_2 - a_2 s_1$$

In order to provide a more succinct definition we introduce a symbol called the *Levi-Civita density*, u_{ijk}, defined by u_{ijk} equal to 1 when i, j, k are all unequal and in the order $1231231\ldots$, equal to -1 when i, j, k are all unequal and in the order $2132132\ldots$, and equal to 0 if any two of i, j, k are equal. Then if $\mathbf{B} = \mathbf{A} \times \mathbf{S}$,

$$b_i = u_{ijk} a_j s_k \tag{1.5}$$

$\mathbf{A} \times \mathbf{S}$ is a vector perpendicular to both $\mathbf{A}$ and $\mathbf{S}$, and whose magnitude is $AS \sin \theta$ where θ is the angle between $\mathbf{A}$ and $\mathbf{S}$, the positive direction being taken from $\mathbf{A}$ to $\mathbf{S}$. The operation is anticommutative and

$$\mathbf{A} \times \mathbf{S} = -\mathbf{S} \times \mathbf{A}$$

We may form multiple vector products such as

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}), \qquad (\mathbf{A} \times \mathbf{B}) \times \mathbf{C}$$

and so on, but it is necessary to use the parentheses to avoid ambiguity since

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) \neq (\mathbf{A} \times \mathbf{B}) \times \mathbf{C}$$

Differentiation of a vector with respect to a scalar. Each of the components of a vector may be a function of some independent scalar variable—for

example, time. In this case, if s_i are components of **S**, then

$$\frac{d\mathbf{S}}{dt}$$

is defined as a vector whose components are

$$\frac{ds_i}{dt}$$

1.4 Physical significance of vectors

Nearly all physical quantities that have both magnitude and direction can be represented by vectors. The only exception that comes immediately to mind is finite angular displacement. The reader may verify this exception by considering the object shown in Fig. 1.4. If this object is rotated through 90° successively about each of the three axes, its final orientation depends upon the sequence in which the axes are taken. This violates the commutative law of addition, hence angular displacement is not a vector. A partial list of physical and geometric quantities which can be represented as vectors includes force, displacement, velocity, acceleration, current densities (this applies to the flow of *anything*, not just electricity), magnetic field intensity, plane finite areas, and infinitesimal areas.

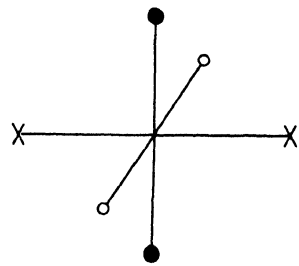


Figure 1.4 This drawing shows, that finite rotation is not a vector. Successive rotation through 90° about each of the three axes will produce different results, depending upon the order in which the axes are chosen.

Much of the material in the first three sections must have seemed rather abstract to the reader; perhaps a greater physical significance can be obtained by considering an example which illustrates some of the various operations defined in Sec. 1.3.

Example 1.1. A charged particle (mass = m, charge = ϵ) is located in a magnetic field of intensity **B** and an electric field of intensity **E**. The total force exerted by these upon the particle is

$$\mathbf{F} = \epsilon(\mathbf{E} + \mathbf{B} \times \mathbf{v})$$

where **v** is the velocity of the particle. According to Newton's law of motion

$$\mathbf{F} = \frac{d}{dt}(m\mathbf{v})$$

If, in a given coordinate system, the components of B and E are

$$b_1 = -100, \qquad b_2 = 0, \qquad b_3 = 0$$
$$e_1 = 50t, \qquad e_2 = 25, \qquad e_3 = -40$$
$$(t = \text{time})$$

set up the differential equations of motion of the particle.

The complete differential equation in vector form is

$$m\frac{d\mathbf{v}}{dt} = \epsilon(\mathbf{E} + \mathbf{B} \times \mathbf{v})$$

The components of the vectors are

$$m\frac{dv_i}{dt} = \epsilon[E_i + u_{ijk}b_j v_k]$$

which represents in more brief form the three simultaneous differential equations

$$m\frac{dv_1}{dt} = 50\epsilon t$$
$$m\frac{dv_2}{dt} = (25 - 100v_3)\epsilon$$
$$m\frac{dv_3}{dt} = -(40 + 100v_2)\epsilon$$

1.5 Unit vectors

Those readers who have taken a course in vector analysis may be somewhat more familiar with a notational system in which unit vectors are designated along a set of coordinate axes, and in which the vectors themselves are represented in terms of these axes. This notational system is not nearly so concise as that which we have developed, but often what it lacks in notational convenience is compensated for by greater clarity. We shall describe the system here, but our subsequent usage of it will be limited to a few cases where it is believed that the other system would tend to be confusing.

Having selected a set of coordinate axes x_i, we designate a set of vectors of unit magnitude $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$ such that $\mathbf{C}_1$ has the direction of increasing x_1, $\mathbf{C}_2$ the direction of increasing x_2, etc. In this notational system any vector $\mathbf{S}$ having components along the axes whose magnitudes are s_1, s_2, s_3 is regarded as the sum of the components:

$$\mathbf{S} = s_1\mathbf{C}_1 + s_2\mathbf{C}_2 + s_3\mathbf{C}_3 \tag{1.6}$$

The operation $\mathbf{A} + \mathbf{S}$ means that

$$\mathbf{A} + \mathbf{S} = (a_1 + s_1)\mathbf{C}_1 + (a_2 + s_2)\mathbf{C}_2 + (a_3 + s_3)\mathbf{C}_3 \tag{1.7}$$

The *dot product* is evaluated from the set of identities

$$\mathbf{C}_i \cdot \mathbf{C}_j = 0, \qquad i \neq j$$
$$= 1, \qquad i = j$$

Thus

$$\mathbf{A} \cdot \mathbf{S} = (a_1\mathbf{C}_1 + a_2\mathbf{C}_2 + a_3\mathbf{C}_3) \cdot (s_1\mathbf{C}_1 + s_2\mathbf{C}_2 + s_3\mathbf{C}_3) = a_1s_1 + a_2s_2 + a_3s_3$$

The vector product may be evaluated from the set

$$\mathbf{C}_1 \times \mathbf{C}_2 = \mathbf{C}_3 \qquad \mathbf{C}_2 \times \mathbf{C}_3 = \mathbf{C}_1$$
$$\mathbf{C}_3 \times \mathbf{C}_1 = \mathbf{C}_2 \qquad \mathbf{C}_i \times \mathbf{C}_j = -\mathbf{C}_j \times \mathbf{C}_i$$
$$\therefore \mathbf{C}_i \times \mathbf{C}_{(i)} = 0 \qquad (i = 1, 2, 3)$$

Thus

$$\mathbf{A} \times \mathbf{B} = (a_1\mathbf{C}_1 + a_2\mathbf{C}_2 + a_3\mathbf{C}_3) \times (b_1\mathbf{C}_1 + b_2\mathbf{C}_2 + b_3\mathbf{C}_3) \tag{1.8}$$
$$= \begin{vmatrix} \mathbf{C}_1 & \mathbf{C}_2 & \mathbf{C}_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}$$

In a like manner, we may differentiate with respect to a scalar:

$$\frac{d\mathbf{A}}{dt} = \frac{da_1}{dt}\mathbf{C}_1 + \frac{da_2}{dt}\mathbf{C}_2 + \frac{da_3}{dt}\mathbf{C}_3 + a_1\frac{d\mathbf{C}_1}{dt} + a_2\frac{d\mathbf{C}_2}{dt} + a_3\frac{d\mathbf{C}_3}{dt}$$

Notice that the last three terms are zero unless the coordinate axes themselves are changing their orientation as time goes by. Often in dynamics it is useful to use rotating coordinates, and this notational system is preferred because it is capable of dealing explicitly with rotating coordinates. In the study of properties of materials we do not have rotating coordinates, so this notational system does not have the aforementioned advantage.

1.6 Scalar and vector fields

If there is some scalar quantity which is functionally related to the variables x_i, then a *scalar field* is said to exist everywhere that this functional relationship exists. Typical examples of scalar fields are the gravitational potential around a massive object, the pressure distribution in a fluid, and the temperature distribution in some material. In each of these cases, a specification of a set of values for the variables x_i will provide a unique value of some scalar quantity.

Vector fields are defined in an analogous manner. Let a_j be a set of linearly independent components of a vector **A**. If each of the components a_j is a function of the variables x_i, then **A** will be a vector function of the space defined by the set x_i. In general, both the magnitude and the direction of **A** will vary from point to point in the space, and specification of values for the variables x_i will result in a unique vector **A**. Any region of a space in which these conditions are met is said to contain a vector field. There are numerous physical examples of vector fields, such as the force field due to gravitation around a massive object. If any single type of vector field can be thought of as having pre-eminent importance in the study of properties of materials, it is that type of field which we call a *current density*. Imagine

that there is a flow of heat, mass, or electricity through some surface. Let $d\mathbf{\sigma}$ be an elementary vector normal to this surface at a point on the surface, and let the magnitude, $d\sigma$, represent an element of area around this point. Then the rate of flow through this elementary area is $\mathbf{J} \cdot d\mathbf{\sigma}$ where $\mathbf{J}$ is a vector function of the coordinates called the *current density*.

In much of the discussion that follows, it will be necessary to consider the fields under discussion to be continuous and continuously differentiable. Unless it is specified otherwise, this assumption will be taken for granted. We also stipulate that the conclusions which we draw will normally be valid only in some bounded region. Some of the general conclusions which we shall establish shortly in regard to scalar and vector fields would be valid within either material, but would not be valid across the interface between different materials.

1.7 The gradient

If the set x_i defines a Cartesian coordinate system and if a scalar field

$$\phi = \phi(x_i) \tag{1.9}$$

exists in some region of this space, we may derive a vector field whose components are

$$\frac{\partial \phi}{\partial x_i}$$

If we evaluate the components of this vector at any point X_i, this vector is called the *gradient* of ϕ at that point.

In terms of a set of unit vectors, $\mathbf{C}_i$, parallel to the coordinate axes x_i,

$$\mathbf{grad}\ \phi = \frac{\mathbf{C}_1 \partial \phi}{\partial x_1} + \frac{\mathbf{C}_2 \partial \phi}{\partial x_2} + \ldots + \frac{\mathbf{C}_n \partial \phi}{\partial x_n} = \frac{\mathbf{C}_i \partial \phi}{\partial x_i} \tag{1.10}$$

We may provide a geometric interpretation of the gradient for a two-dimensional space by constructing the surface shown in Fig. 1.5, where the magnitude of the scalar ϕ is plotted along a vertical axis as a function of the variables x_1 and x_2. The gradient of ϕ at a point X_1, X_2 is a vector whose magnitude is equal to the slope of the surface at the point and is normal to the surface at the same point.

The gradient is an important and highly practical concept. Its chief utility is in dealing with forces or flows caused by the gradients of scalar physical quantities. Examples are pressure, temperature, concentration, and electric and gravitational potentials. If whatever moves under the influence of these gradients encounters a resistance, then in the steady state, the flow current density will be proportional to the gradient. In the cases of greatest interest in this text these proportionality factors will be properties of the material in which the flow occurs. In general, these will be tensors of rank two, but for the present we may treat them as scalars.

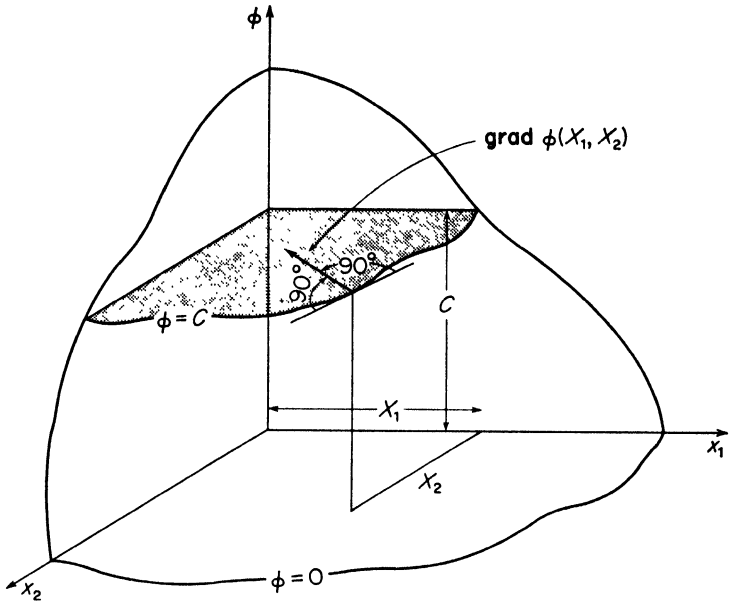


Figure 1.5 Geometrical representation of the gradient of a function of two space coordinates.

Using these ideas then, we can write a number of relationships involving different physical quantities which are all of essentially the same form:

$$-\,\mathbf{grad}\ T = {}_qK_q\mathbf{J}, \qquad T = \text{temperature}$$
$${}_qK = \text{thermal conductivity}$$
$${}_q\mathbf{J} = \text{heat current density}$$
$$-\,\mathbf{grad}\ V = {}_eK_e\mathbf{J}, \qquad V = \text{electric potential}$$
$${}_eK = \text{electrical conductivity}$$
$${}_e\mathbf{J} = \text{electrical current density}$$
$$-\,\mathbf{grad}\ C = {}_mK_m\mathbf{J}, \qquad C = \text{concentration}$$
$${}_mK = \text{diffusivity}$$
$${}_m\mathbf{J} = \text{mass current density}$$

The reader should be warned that these equations apply only to the steady state. Transient problems are not much more complex to formulate but we need to postpone a consideration of these until after the next section has been developed.

Since every scalar field having the required continuity will possess a gradient at every point, each scalar field must have an associated vector field. Furthermore, these are point functions of the coordinates, and in general the gradient vector will vary from point to point in space subject only to continuity restrictions.

In a case where the medium permits no flow, a gradient will still produce a driving force. For example, free charges embedded in a plastic are not very mobile, and little or no steady-state flow of charge will result from an electric potential gradient. Nevertheless, they will be subject to a force which is proportional to the gradient.

Example 1.2. The electric potential in a particular region is given by

$$V = \frac{1}{(x^2 + y^2 + z^2)^{1/2}}$$

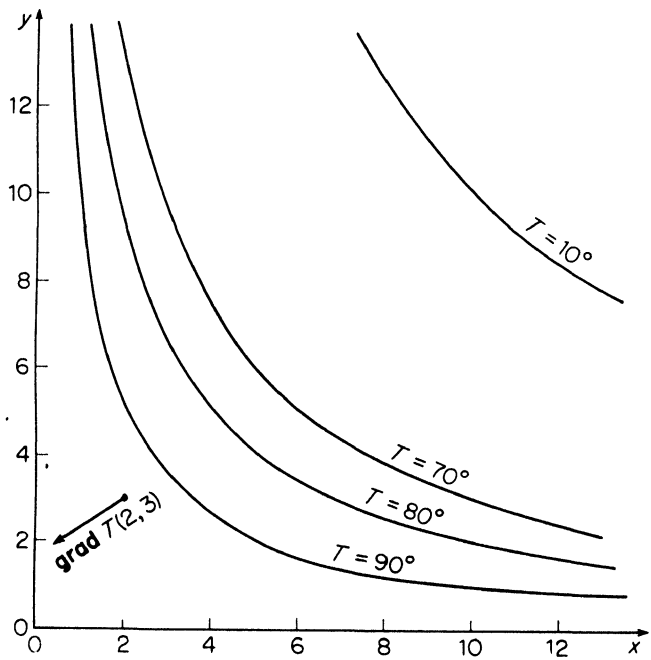
If the electrical conductivity is ${}_eK$, compute the current density at the point (3, 2, 2).

Answer:

$${}_eK_e\mathbf{J} = -\,\mathbf{grad}\, V = \frac{(x\mathbf{C}_1 + y\mathbf{C}_2 + z\mathbf{C}_3)}{(x^2 + y^2 + z^2)^{3/2}}$$

$${}_e\mathbf{J} = \frac{3\mathbf{C}_1 + 2\mathbf{C}_2 + 2\mathbf{C}_3}{{}_eK(17)^{3/2}}$$

Example 1.3. The temperature distribution in a flat plate is given by $T = 100 - xy$. The figure shows the plate with 90°, 80°, 70°, and 10° isotherms



drawn in. The gradient vector at any point on an isotherm will be perpendicular to the isotherm and directed toward a region of higher temperature. In general

$$\mathbf{grad}\, T = -y\mathbf{C}_1 - x\mathbf{C}_2$$

At a particular point, say $x = 2, y = 3$,

$$\mathbf{grad}\ T_{(2,3)} = -3\mathbf{C}_1 - 2\mathbf{C}_2$$

1.8 Line, surface, and volume integrals

We have defined the differentiation of a vector in Sec. 1.3, but so far we have not used any integrals in which vectors were involved. There are three types of integral operations involving vectors which, although not directly useful in themselves, are necessary for the development of some very important theorems. These are line, surface, and volume integrals.

The definition of a *line integral* is embodied in the equation

$$\int_c \mathbf{A} \cdot d\mathbf{s} = \lim_{\Delta s_i \to 0} \sum \mathbf{A}(x_i) \cdot \Delta \mathbf{s}_i \tag{1.11}$$

where c represents some specific curve and $d\mathbf{s}$ is an elementary displacement along the curve. The symbol $\mathbf{A}(x_i)$ means that $\mathbf{A}$ is a point function of the coordinates and $\mathbf{A}(x_i)$ is the value of $\mathbf{A}$ at some point along the curve where the displacement $\Delta \mathbf{s}_i$ includes the points whose coordinates are x_i. If $\mathbf{A}(x_i)$ were a force field, then the line integral of $\mathbf{A}$ along a curve would represent the work done upon a particle by the force field as it moved along the curve.

If $\mathbf{A}$ is the gradient of a scalar $\phi(x_1, x_2, x_3)$, we could write

$$\mathbf{A} = \frac{\mathbf{C}_1 \partial\phi}{\partial x_1} + \frac{\mathbf{C}_2 \partial\phi}{\partial x_2} + \frac{\mathbf{C}_3 \partial\phi}{\partial x_3}$$

and

$$d\mathbf{s} = \mathbf{C}_1\, dx_1 + \mathbf{C}_2\, dx_2 + \mathbf{C}_3\, dx_3$$

Then

$$\int_c \mathbf{A} \cdot d\mathbf{s} = \int_c \frac{\partial\phi}{\partial x_1} dx_1 + \frac{\partial\phi}{\partial x_2} dx_2 + \frac{\partial\phi}{\partial x_3} dx_3$$

In this case the integrand is a perfect differential, and between any two points the value of the line integral will be independent of the path, provided the region is simply connected. This point will frequently be useful in evaluating line integrals when they involve vectors derivable from the gradient of a scalar.

Surface integrals are defined by the equation

$$\int_\Sigma \mathbf{A} \cdot d\boldsymbol{\sigma} = \lim_{\Delta\sigma_i \to 0} \sum \mathbf{A}(x_i) \cdot \mathbf{n}\, \Delta\sigma_i \tag{1.12}$$

The meaning of this definition can be clarified by considering an arbitrary, finite, simple closed surface Σ, enclosing a region τ, and choosing a point on the surface whose coordinates are x_i. If a vector field $\mathbf{A}$ exists in the region containing the surface and $\Delta\sigma_i$ is a small section of the surface containing the point x_i, then the outflow of $\mathbf{A}$ through this portion of the surface will be equal to $\mathbf{A} \cdot \mathbf{n}\, \Delta\sigma_i$, where $\mathbf{A}$ is the average value of $\mathbf{A}$ at points inside the

area $\Delta\sigma_i$, and $\mathbf{n}$ is a unit vector normal to the surface directed outward from the region τ.

If the surface Σ is closed, then $\int_\Sigma \mathbf{A} \cdot d\boldsymbol{\sigma}$ represents the net outflow of $\mathbf{A}$ from the region enclosed by Σ. For example, if $\mathbf{A}$ represents a fluid current density, then $\int_\Sigma \mathbf{A} \cdot d\boldsymbol{\sigma}$ will be the net rate of outflow of fluid from the region.

The *volume integral* of a scalar ϕ in τ is defined by

$$\int_\tau \phi \, \Delta\tau = \lim_{\Delta\tau_i \to 0} \sum \phi_i \, \Delta\tau_i \tag{1.13}$$

if ϕ is a suitably continuous function of the coordinates in the region τ. For example, if C represents the concentration of some material in τ, and if there were no generation of this material in τ, the net outflow of material from the region could be expressed by the integral

$$-\int_\tau \frac{\partial C}{\partial t} \, d\tau$$

This concept can be extended to include vector fields; if a_j are the components of a vector $\mathbf{A}$ and if each a_j is a function of the coordinates x_i, then

$$\int_\tau \mathbf{A} \, d\tau = \mathbf{C}_1 \int_\tau a_1 \, d\tau + \mathbf{C}_2 \int_\tau a_2 \, d\tau + \ldots + \mathbf{C}_n \int_\tau a_n \, d\tau \tag{1.14}$$

1.9 The vector operator and the divergence of a vector

The operation performed upon ϕ in Eq. (1.10) occurs frequently and this operation

$$\mathbf{C}_1 \frac{\partial}{\partial x_1} + \ldots + \mathbf{C}_n \frac{\partial}{\partial x_n}$$

is given the symbol ∇. This is to be regarded as a vector operator which may operate upon a scalar or may form a dot, cross, or dyadic product with a vector. Thus

$$\mathbf{grad}\ \phi = \nabla\phi \tag{1.15}$$

$$\nabla \cdot \mathbf{A} = \frac{\partial a_i}{\partial x_i}, \qquad \nabla \times \mathbf{A} = \mathbf{B}, \qquad b_k = u_{ijk} \frac{\partial a_j}{\partial x_i}$$

where u_{ijk} is the Levi-Civita density.

$$\nabla \cdot u\mathbf{A} = \nabla u \cdot \mathbf{A} + u\nabla \cdot \mathbf{A}$$

We may encounter any one of these operations in problems later, but of these, the operation $\nabla \cdot \mathbf{A}$ is of particular importance. This is called the *divergence* of $\mathbf{A}$, and in three-dimensional space

$$\nabla \cdot \mathbf{A} = \frac{\partial a_1}{\partial x_1} + \frac{\partial a_2}{\partial x_2} + \frac{\partial a_3}{\partial x_3}$$

This, however, is not the defining equation for div **A**. The *divergence* of **A** is the net outflow of **A** per unit volume at a point and is defined by the equation

$$\text{Div } \mathbf{A} = \lim_{\substack{\tau \to 0 \\ \Sigma \to 0}} \frac{\int_\Sigma \mathbf{A} \cdot d\boldsymbol{\sigma}}{\tau} \tag{1.16}$$

where Σ is a surface surrounding the volume τ which includes a point x_i, and **A** is the value of **A** inside τ or on Σ.

The reader is invited to prove the relationship Div $\mathbf{A} = \nabla \cdot \mathbf{A}$ for himself.

Equation (1.16) may be written in the equivalent form

$$\int_\tau \text{Div } \mathbf{A}\, d\tau = \int_\Sigma \mathbf{A} \cdot d\boldsymbol{\sigma} \tag{1.17}$$

This is called the *divergence theorem;* it plays a very important role in continuum mechanics.

By definition the divergence of a vector is the net outflow of that vector per unit volume at a point. If, for example, the vector were representing rate of flow of heat per unit area and there were no heat sources in the region under consideration, then the system would be described by the differential equation $\nabla \cdot \nabla T = 0$. On the other hand, if there were heat sources present (examples ranging from compost heaps to fuel rods in a reactor spring to mind), then the divergence at a point would be equal to the strength of the source at that point. We shall study this point in greater detail in the next section.

1.10 Applications

In the previous sections, a definite attempt was made to keep the mathematics tied to physical concepts. However, as isolated concepts the gradient and divergence have only a limited usefulness. It is when they are used in conjunction with each other and with the divergence theorem that they provide powerful tools for setting up partial differential equations describing physical situations.

In illustrating the use of these concepts, we shall also wish to speak of sources and sinks. These terms are almost self-explanatory, but to be sure that there is no misunderstanding, we shall consider a *source* to be a point at which something is generated. For example, in the esterification of an acid and an alcohol, water will be generated at any point in the system at the rate at which the acid and the alcohol are combining at that point. Another example would be the force field around a charged particle. The force can be thought of as continually flowing outward from a source, i.e. the charged particle. A *sink* is just the opposite, i.e. it is a point at which something disappears.

We shall also wish to distinguish between steady state and transient situations. For example, if we were to heat one side of an insulating pad at a steady rate, initially the temperature distribution in the pad would be time dependent. During this period, the system is in a transient state, but after a certain period of time has elapsed, the temperature distribution will have become a function of the coordinates but not of time. Then the system will have attained a steady-state condition.

Let us first consider a rather general situation; let $\mathbf{A}(x_i)$ be a continuous vector field in some finite region bounded by a simple closed surface and let $d\mathbf{A}(x_i)/dt$ vanish everywhere in the region. We consider an arbitrary subregion τ bounded by a surface Σ. We specify that there are no sources or sinks in any such subregion. As a consequence, the outflow of $\mathbf{A}(x_i)$ from this subregion will equal the inflow of $\mathbf{A}(x_i)$, or in other words the *net* outflow of $\mathbf{A}(x_i)$ from this subregion will be zero:

$$\int_\Sigma \mathbf{A} \cdot d\mathbf{\sigma} = 0$$

But by the divergence theorem

$$\int_\Sigma \mathbf{A} \cdot d\mathbf{\sigma} = \int_\tau \nabla \cdot \mathbf{A}\, d\tau = 0$$

But since τ is any arbitrary subregion,

$$\nabla \cdot \mathbf{A} = 0$$

This differential equation will then be valid throughout the region under discussion. Often, however, this is not a useful form for applying boundary conditions, but if $\mathbf{A}(x_i)$ is the gradient of some scalar ϕ, on which we can apply boundary conditions, then

$$\nabla \cdot \mathbf{A} = \nabla \cdot \nabla\phi = 0$$

Now let us apply these ideas to a specific case. Consider an isotropic material containing no heat sources or sinks and which has reached a steady state with respect to a set of temperature conditions applied to its boundaries. Let K be the thermal conductivity of the material and let $\mathbf{J}$ represent the entropy current density anywhere in the material. This is related to the temperature gradient at any point by

$$\mathbf{J} = -K\nabla T \qquad (T = \text{temperature})$$

and by virtue of the absence of heat sources or sinks inside the boundaries of the material, and the condition that a steady state has been attained,

$$\nabla \cdot \mathbf{J} = 0$$

Thus

$$\nabla \cdot K\nabla T = \nabla K \cdot \nabla T + K\nabla \cdot \nabla T = 0$$

The solution of this differential equation for any given set of boundary conditions will give the temperature distribution anywhere in the material. From this the entropy current density can be calculated at any point in the material. The thermal conductivity K would not be a function of the coordinates in a homogeneous material except for the fact that K may be temperature dependent. If $K = f(T)$, then

$$\nabla K = f(T)\left\{\mathbf{C}_1\frac{\partial T}{\partial x_1} + \mathbf{C}_2\frac{\partial T}{\partial x_2} + \mathbf{C}_3\frac{\partial T}{\partial x_3}\right\}$$

and

$$\nabla \cdot K\nabla T = f'(T)\frac{\partial T}{\partial x_i}\frac{\partial T}{\partial x_i} + \frac{K\,\partial^2 T}{\partial x_i\,\partial x_i} = 0$$

In a similar manner, if a gradient of a scalar produces a force field and no force sources are present within the region of definition, then the net outflow of force per unit volume will be zero throughout the region under consideration. For example, if V is a gravitational (or electrical) potential, then in a region in which there are no masses (or free electric charges) present, $\nabla \cdot \mathbf{F} = 0$. If $\mathbf{F} = -\nabla V$, then $\nabla \cdot \nabla V = 0$.

When a region is heterogeneous, then separate differential equations must be set up for each homogeneous subregion, and the solutions must be matched at the boundaries of the subregions. As an example of this, consider the flow of heat through a region containing materials A and B, where A has a higher thermal conductivity than B. If isothermal lines are plotted as shown in Fig. 1.6, then at the interface going from B to A the isotherms will be bent backwards because of the higher thermal conductivity of A. This produces a discontinuity in the derivatives of ∇T. Inside region A, the equation

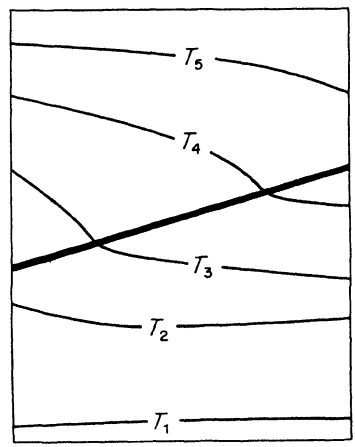


Figure 1.6 Refraction of isotherms at the interface between two materials having different thermal conductivity. The upper region has the higher thermal conductivity.

$$\nabla \cdot K_A\nabla T = 0$$

applies, and inside region B, the equation

$$\nabla \cdot K_B\nabla T = 0$$

applies. Solutions must be chosen for matching boundary conditions at the interfaces.

When sources are present, the net outflow through a small enclosed subregion will be proportional to the total number of sources in it. This can be

represented by

$$\text{net outflow} = \int_\tau c\rho(x_i)\, d\tau = \int_\tau \mathbf{J} \cdot d\boldsymbol{\delta}$$

where $\rho(x_i)$ is the distribution or density of sources per unit volume and c is the rate of production per source. By the divergence theorem, we may write

$$\int_\Sigma \mathbf{J} \cdot d\boldsymbol{\delta} = \int_\tau \nabla \cdot \mathbf{J}\, d\tau$$

where Σ bounds any arbitrary subregion. Consequently

$$\nabla \cdot \mathbf{J} = c\rho(x_i)$$

Now to consider a more concrete example, suppose these are heat sources. Then, since $\mathbf{J}$ equals $-K\,\nabla T$,

$$\nabla \cdot (K\,\nabla T) = -c\rho(x_i)$$

or

$$K\nabla^2 T + \nabla K \cdot \nabla T = -c\rho(x_i)$$

where K may be a function of temperature and ρ is a function of the coordinates.

As a further illustration of the use of the divergence theorem, let us derive the differential equation which describes the electric potential in a region containing free charges. We may divide this region into infinitesimal subregions, each of which contains only one charged particle. We may then compute the net outflow of the force field from each particle. Since the subregions are infinitesimal we may choose their shape at our convenience; a spherical subregion of radius r with the particle at the center will serve nicely. The potential at the surface of the sphere is q/Kr and the net outflow of the force field is

$$\nabla\left(\frac{q}{Kr}\right) \cdot d\boldsymbol{\delta}$$

In spherical coordinates, $d\boldsymbol{\delta} = \mathbf{r}_1 r^2 \sin\theta\, d\theta\, d\phi$, and

$$\frac{q\nabla}{K}\left(\frac{1}{r}\right) = \frac{-q\mathbf{r}_1}{Kr^2}$$

where $\mathbf{r}_1$ is a unit vector directed outward from the surface of the sphere.

Consequently, $8\int_0^{\pi/2}\int_0^{\pi/2} \frac{qr^2}{Kr^2} \sin\theta\, d\theta\, d\phi\, \mathbf{r}_1 \cdot \mathbf{r}_1 = \frac{-4\pi q}{K}$

Taking a larger subregion containing a number of charges, the net outflow per unit volume will be equal to the charge density times the volume times the outflow per unit charge divided by the volume. Hence

$$\nabla \cdot \nabla V = \frac{-4\pi\rho}{K}$$

This is called *Gauss' theorem.* The 4π is incorporated into the permittivity in the rationalized system of units.

Transient systems. Even though no true sources are present when a system is in a transient state, there may be either a rate of accumulation or a rate of discharge from different points in a region. To cite a specific example, if water vapor is diffusing through a membrane, it is a logical necessity that there be water *inside* the membrane. If the membrane were initially free of any dissolved water, then when diffusion started, a transient accumulation of water inside the membrane would be superimposed upon the flow of water through the membrane. This phenomenon can be generalized into the so-called *continuity equation.* If $\mathbf{J}$ is the current density of whatever is flowing and $c(x_i)$ is the concentration at a point x_i, then the net outflow through the boundaries of a subregion Σ is the negative of the rate of accumulation in the subregion. Thus, using the divergence theorem again,

$$\int_{\Sigma} \mathbf{J} \cdot d\mathbf{\sigma} = -\int \frac{\partial c(x_i)}{\partial t}\, d\tau = -\int_{\tau} \nabla \cdot \mathbf{J}\, d\tau$$

Hence

$$\nabla \cdot \mathbf{J} = \frac{-\partial c(x_i)}{\partial t}$$

Example 1.4. If $\mathbf{J}$ represents rate of flow of entropy per unit area and a transient state is under consideration, then the term $c(x_i)$ represents entropy per unit volume at a point. If ρ is the mass density of the medium through which the heat is flowing, then

$$c(x_i) = \int_0^T \frac{\rho c_p\, dT}{T}$$

Then the continuity equation becomes

$$-\nabla \cdot (K \nabla T) = \frac{-\partial c}{\partial t}$$

Example 1.5. If $\mathbf{J}$ represents the current density of a compressible fluid, then $c(x_i)$ is equivalent to the density of the fluid and $\mathbf{J} = \rho\mathbf{v}$ where $\mathbf{v}$ is the velocity. Thus the velocity field will be characterized by the differential equation

$$\nabla \cdot (\rho\mathbf{v}) = \frac{-\partial c}{\partial t}$$

1.11 Summary

The previous section should have served to indicate the use of vector field theory in obtaining the differential equations describing the state of a continuous medium. Our use of vector fields in the study of solid materials is somewhat more restricted than this, since the whole of our effort shall be

directed toward characterizing the way a material responds to its environment. In this sense, if the reader is able to understand all of the implications of an equation such as

$$\nabla \cdot \mathbf{J} = -K \nabla T$$

we shall have accomplished our objective, since this provides a precise definition of what is meant by the thermal conductivity K. Naturally the objective of this chapter has not been to provide a precise definition of thermal conductivity; instead, this should be regarded as a part of the mathematical apparatus that we shall require later in defining and characterizing material properties.

PROBLEMS

1.1 (a) Expand the set, $a_{ij}b_j = 0$, when $i, j = 1, 2$.
(b) Expand the set, $a_{ij}b_{jk} = \delta_{ik}$, when $i, j, k = 1, 2, 3$.
(c) Expand the set, $a_{ijk}b_k = 0$, when $i, j, k = 1, 2$.
(d) Expand the set, $a_{(ij)}b_{(j)}$, when $i, j = 1, 2$.
(e) Expand the set, $T'_{ijk} = a_{il}a_{jm}a_{kn}T_{lmn}$, when $i, j, k, l, m, n = 1, 2$.

1.2 Express the following sets of equations in terms of a single equation with letter subscripts on each factor. Employ the Kronecker delta or the Levi-Civita density where applicable.
(a) $a_1b_1 + a_2b_2 + a_3b_3 = c$
(b) $a_{11}b_1 + a_{12}b_2 + a_{13}b_3 = c_1$
$a_{21}b_1 + a_{22}b_2 + a_{23}b_3 = c_2$
$a_{31}b_1 + a_{32}b_2 + a_{33}b_3 = c_3$
(c) $a_{11}b_{11} + a_{12}b_{21} = 1$
$a_{21}b_{12} + a_{22}b_{22} = 1$
$a_{11}b_{12} + a_{12}b_{22} = 0$
(d) $a_2b_3 - a_3b_2 = c_1$
$a_3b_1 - a_1b_3 = c_2$
$a_1b_2 - a_2b_1 = c_3$
(e) $a_{11}b_1 + a_{22}b_2 + a_{33}b_3 = c$

1.3 Let $A_1 = 2$, $A_2 = -1$, $A_3 = 3$ be a set of components of the vector **A** along a particular set of Cartesian coordinate axes and let $B_1 = -1$, $B_2 = 4$, $B_3 = 0$ be a corresponding set of components of the vector **B**.
(a) Evaluate $\mathbf{A} \cdot \mathbf{B}$ and the components of $\mathbf{C} = \mathbf{A} \times \mathbf{B}$ and $\mathbf{D} = \mathbf{B} \times \mathbf{A}$ along the axes.
(b) Evaluate the cosine of the angle between **A** and **B**.
(c) Evaluate the component of **A** in the direction of the unit vector **n** whose components along the coordinate axes are all positive and equal.
(d) Evaluate the components of **A** in the direction of the unit vector **n** whose components are found from $n_1 = -n_2$, $n_3 = 0$.

1.4 A plane may be represented by an equation of the form

$$ax_1 + bx_2 + cx_3 + d = 0.$$

Show that a unit vector normal to the plane will have the components

$$n_1 = \frac{a}{(a^2 + b^2 + c^2)^{1/2}} \qquad n_2 = \frac{b}{(a^2 + b^2 + c^2)^{1/2}} \qquad n_3 = \frac{c}{(a^2 + b^2 + c^2)}$$

Hint: Explore the consequences of taking the dot product of any vector lying in the plane and the unit normal.

1.5 If $\phi = x_1^2 x_2 x_3$, evaluate $\nabla \cdot \nabla\phi$ and the components of $\nabla\phi$ at the point (1, 1, 1).

1.6 If, in some region, the components of a vector **A** are functionally dependent upon the coordinates according to the equations

$$A_1 = 3x_1 + 2 \qquad A_2 = x_1 + x_2^2 + x_3 \qquad A_3 = 2x_1 + x_3,$$

find the net outflow of **A** per unit volume at the point (2, 2, 2).

1.7 Let $\phi = x_1^2 + x_2^2 + x_3^2$ and let $\mathbf{A} = \nabla\phi$. Evaluate the line integral $\int_c \mathbf{A} \cdot d\mathbf{S}$ along the curve $x_1^2 + x_2 = 25$ from the point (5, 0, 0) to the point (2, 5, 0).

1.8 The rate at which neutrons are absorbed in a unit volume of a material is equal to the product of the magnitude of the neutron current density, the density of absorbing atoms and their collision cross section. The current density of neutrons at a point is proportional to the concentration gradient of neutrons at that point. The proportionality constant is called the diffusivity of neutrons in the material in question. Set up a differential equation whose solution would give the concentration, c, of neutrons as a function of time and coordinates in a material having a neutron diffusivity, D, and ϕ, density of absorber atoms/cm^3, each of which has a collision cross section of σ cm^2.

1.9 (a) Let C represent the number of grams of water absorbed per unit volume of a permeable material and let D represent the diffusivity of water. The current density of water at a point, the concentration gradient, and the diffusivity are related through the equation

$$\mathbf{J} = -D\nabla C.$$

Assuming that D is independent of C, derive the equation

$$\frac{\partial C}{\partial t} = -D\nabla \cdot \nabla C$$

(b) If D were a function of C, what differential equation would be applicable to the system?

2

MATRIX ALGEBRA

Matrix algebra is designed to facilitate handling sets of equations, sets of functions, or sets of components. For our purposes it is primarily an operational and notational tool which permits the user to express quantities of information in a compact form without the labor of writing it in extenso. This chapter is not intended as an exhaustive treatment of the topic, but is devoted to the pragmatic objective of developing only those matrix operations which will be useful later.

There are alternate ways of characterizing material properties which bear an inverse relationship to one another. This inverse relationship is not a simple reciprocal, but involves other properties as well and can best be stated in terms of the inverse of a matrix of related properties. In a quite different context, we shall also need certain matrix operations to show the relationship between crystal structure and anisotropy of properties of single crystals. These two examples are typical of the subsequent use of matrices which we anticipate now.

2.1 Basic definitions and operations

A *matrix* is nothing more or less than an *array of symbols arranged in rows and columns.* Normally we use the same symbol throughout the array and designate the row and column by subscripts. Written in extenso, a general $n \times m$ matrix would have the form

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

When $n = 1$, the matrix will be a *row matrix*, and when $m = 1$, it is a *column matrix*. The range of the first subscript is all of the integers from 1 to n and that of the second is all of the integers from 1 to m.

A general element in the ith row and the jth column is represented by the symbol a_{ij}. The entire matrix may be represented by either a general symbol surrounded by parentheses or a capital letter: (a_{ij}) or $\mathbf{A}$. The range of i and j is usually understood from the context in which it is used, but it may be explicitly stated when it is necessary to do so for clarity. For example, if the range were not known from context, we might write

$$(a_{ij}), \qquad \begin{matrix} i = 1, 2 \\ j = 1, 2, 3 \end{matrix}$$

Matrices in which m equals n are called *square matrices*, and square matrices in which a_{ij} equals a_{ji} are called *symmetric matrices*. Matrices in which the elements a_{ij} equal $-a_{ji}$ are called *skew-symmetric*. The diagonal elements of a skew-symmetric matrix must be zero, since each of these would be equal to its own negative.

Matrices having the same number of rows and columns may be added. The operation

$$(b_{ij}) = (a_{ij}) + (c_{ij}), \qquad \begin{matrix} i = 1, 2, \ldots, n \\ j = 1, 2, \ldots, m \end{matrix}$$

means that the element b_{ij} is equal to a_{ij} plus c_{ij}. For example,

$$b_{32} = a_{32} + c_{32}, \qquad b_{41} = a_{41} + c_{41}$$

Matrices may also be multiplied, but only under restricted conditions. The operation is not commutative and is defined only when the first member of the pair being multiplied has the same number of columns as the second member has rows. If A is an $m \times n$ matrix and B is an $n \times o$ matrix, then the product

$$(c_{ij}) = (a_{ij})(b_{ij})$$

means that each element c_{ij} of the matrix (c_{ij}) is computed from the formula

$$c_{ij} = a_{ik}b_{kj} \tag{2.1}$$

where the repetition of k demands a summation over the range of k. If (a_{ij}) and (b_{ij}) are both 3×3 matrices, then all subscripts have the range of 3. A typical element c_{23} of the matrix (c_{ij}) would be

$$c_{23} = a_{21}b_{13} + a_{22}b_{23} + a_{23}b_{33}$$

Notice that the inner indices of each term are identical with each other, and the outer indices are identical with the corresponding indices of c_{23}.

The following examples should clarify this operation. In multiplying the row matrix (b_{1j}) by a column matrix (a_{i1}), the inner indices of the product $a_{i1}b_{1j}$ have a range of 1, and summation over this range gives only one term. Consequently

$$(a_{i1})(b_{1j}) = (a_{i1}b_{1j})$$

If the range of i and j is 3, the matrix $(a_{i1}b_{1j})$ is a 3×3 matrix. Written in extended form, this becomes

$$\begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} [b_{11} \quad b_{12} \quad b_{13}] = \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{11}b_{13} \\ a_{21}b_{11} & a_{21}b_{12} & a_{21}b_{13} \\ a_{31}b_{11} & a_{31}b_{12} & a_{31}b_{13} \end{bmatrix}$$

If we reverse the operation and evaluate the product $(b_{1j})(a_{i1})$, the result is a matrix having only one row and one column—in other words, a simple scalar quantity.

$$[b_{11} \quad b_{12} \quad b_{13}] \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} = [b_{11}a_{11} + b_{12}a_{21} + b_{13}a_{31}]$$

The product of the 3×3 matrix (a_{ij}) and the column matrix (b_{i1}) gives a column matrix:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} \\ a_{31}b_{11} + a_{32}b_{21} + a_{33}b_{31} \end{bmatrix}$$

However, the operation $(b_{i1})(a_{jk})$ is not defined since the number of rows in (a_{jk}) exceeds the number of columns of (b_{i1}).

We may illustrate the fact that matrix multiplication is not commutative by expanding the products **AB** and **BA**. Taking

$$\mathbf{C} = \mathbf{AB} \qquad \text{and} \qquad \mathbf{D} = \mathbf{BA}$$

and choosing elements in the second row, third column as being typical, we have

$$c_{23} = a_{21}b_{13} + a_{22}b_{23} + a_{23}b_{33}$$

$$d_{23} = b_{21}a_{13} + b_{22}a_{23} + b_{23}a_{33}$$

which illustrates that in general

$$\mathbf{AB} \neq \mathbf{BA}$$

Even though matrix multiplication is not commutative, it is associative, and we may multiply several matrices successively provided the conditions involving the numbers of rows and columns are met:

$$\mathbf{A(BC)} = \mathbf{(AB)C} = \mathbf{ABC}$$

2.2 The unit matrix, the inverse of a matrix, and the transpose matrix

There are certain types of matrices that occur with sufficient frequency that they warrant some individual comment. The first of these that we wish to consider is the *unit matrix* in which all of the diagonal elements are equal to 1 and the nondiagonal terms are zero. In the case of a 3×3 matrix,

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The reader may easily verify that multiplication between this and any other matrix leaves the matrix unchanged. Thus, if $\mathbf{A}$ is a 3×3 matrix,

$$\mathbf{A} = \mathbf{AI}$$

Further, multiplication by the unit matrix is commutative, although this is not generally true of matrix multiplication. Thus

$$\mathbf{A} = \mathbf{AI} = \mathbf{IA}$$

The primary reason for defining the unit matrix is to provide a convenient definition of the inverse of a matrix. If $\mathbf{A}^{-1}$ is the inverse of $\mathbf{A}$, we define the meaning of the term "inverse" by the equation

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

We shall wish to compute the elements of the inverse matrix in terms of the elements of the direct matrix. Let b_{ij} represent an element of $\mathbf{A}^{-1}$ and a_{ij} an element of $\mathbf{A}$. Then

$$(a_{ij})(b_{jk}) = \mathbf{I}$$

but since all the nondiagonal terms of $\mathbf{I}$ are zero,

$$\begin{aligned} a_{ij}b_{jk} &= 1, \qquad i = k \\ &= 0, \qquad i \neq k \end{aligned} \tag{2.2}$$

If we expand this set of simultaneous equations, we obtain

$$\begin{aligned} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} &= 1, & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} &= 0 \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} &= 0, & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} &= 1 \\ a_{31}b_{11} + a_{32}b_{21} + a_{33}b_{31} &= 0, & a_{31}b_{12} + a_{32}b_{22} + a_{33}b_{32} &= 0 \end{aligned}$$

$$\begin{aligned} a_{11}b_{13} + a_{12}b_{23} + a_{13}b_{33} &= 0 \\ a_{21}b_{13} + a_{22}b_{23} + a_{23}b_{33} &= 0 \\ a_{31}b_{13} + a_{32}b_{23} + a_{33}b_{33} &= 1 \end{aligned}$$

This set of equations may be solved by determinants for any b_{ij} in terms of a_{ij}. Thus, for example,

$$b_{23} = \frac{\begin{vmatrix} a_{11} & 0 & a_{13} \\ a_{21} & 0 & a_{23} \\ a_{31} & 1 & a_{33} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}}$$

The denominator is simply the determinant of the elements of **A** and is the same for any element of the matrix $\mathbf{A}^{-1}$. The numerator is called the *cofactor* of the element a_{ij}. This can be formed by replacing each element in the ith column of the determinant of **A** with a zero except for the element in row j which is replaced by one. The inverse of a matrix exists, of course, only if the determinant formed from its elements is not equal to zero. The reader may take some comfort from the fact that it is rarely necessary to actually compute all elements of the inverse of a matrix. Often it suffices to merely symbolize it. In those cases where a computation is needed, it is often only one element of the inverse matrix that is required.

Example 2.1. A set of quantities x_i are related to another set y_i by the matrix equation

$$(x_i) = (a_{ij})(y_i) \qquad (i, j = 1, 2, 3)$$

Express y_2 as a function of the quantities x_i and the coefficients a_{ij}.

Operating upon both sides of the previous equation gives the sequence

$$(a_{ij})^{-1}(x_i) = \mathbf{I}(y_i)$$

Therefore

$$y_2 = b_{21}x_1 + b_{22}x_2 + b_{23}x_3$$

where

$$b_{21} = \frac{\begin{vmatrix} a_{11} & 1 & a_{13} \\ a_{21} & 0 & a_{23} \\ a_{31} & 0 & a_{33} \end{vmatrix}}{\text{Det } \mathbf{A}}, \quad b_{22} = \frac{\begin{vmatrix} a_{11} & 0 & a_{13} \\ a_{21} & 1 & a_{23} \\ a_{31} & 0 & a_{33} \end{vmatrix}}{\text{Det } \mathbf{A}}, \quad b_{23} = \frac{\begin{vmatrix} a_{11} & 0 & a_{13} \\ a_{21} & 0 & a_{23} \\ a_{31} & 1 & a_{33} \end{vmatrix}}{\text{Det } \mathbf{A}}$$

The *transpose* of a matrix is obtained by a simple transposition of rows and columns and is designated by a tilde over the letter representing the matrix, $\tilde{\mathbf{A}}$. If b_{ij} is a general element of $\tilde{\mathbf{A}}$ and a_{ji} is a general element of **A**, then

$$b_{ij} = a_{ji}$$

We shall have occasion in the next section to deal with matrices whose elements are restricted to values permitted by the set of equations

$$\begin{aligned} a_{ij}b_{ik} &= 1, \quad j = k \\ &= 0, \quad j \neq k \end{aligned} \tag{2.3}$$

The inverse of such a matrix is equal to its transpose:

$$\mathbf{A}^{-1} = \tilde{\mathbf{A}}$$

The proof of this statement follows from the fact that b_{ji} equals a_{ij}. If this is the case, then

$$\begin{aligned} a_{ji}b_{ik} &= 1, \quad j = k \\ &= 0, \quad j \neq k \end{aligned}$$

but this is precisely the condition required of the elements of $\mathbf{A}^{-1}$. Therefore $\mathbf{A}^{-1}$ equals $\tilde{\mathbf{A}}$ as long as the elements of $\mathbf{A}$ are restricted by Eqs. (2.3).

2.3 Matrices as rotation operators

In Chaps. 6 and 8 the variation of properties of solids with direction will be correlated with the symmetry elements of the structure of crystals. In treating this topic we shall make extensive use of transformations of the components of vectors, tensors, and symmetry elements from one coordinate system to another. Matrix operators are particularly useful in this context. We may indicate how this works by the following example.

Let x_i and y_i ($i = 1, 2, 3$) each represent a different set of coordinate axes with coincident origins and let a_{ij} be the cosine of the angle between the y_i axis and the x_j axis. If $\mathbf{r}$ is some vector whose components in the x_i coordinate system are X_i and whose components in the y_i coordinate system are Y_i, then the sets X_i and Y_i are related through the set of equations

$$Y_i = a_{ij}X_j$$

This could be expressed by the matrix equation

$$\mathbf{Y} = \mathbf{AX} \tag{2.4}$$

where a_{ij} are the elements of $\mathbf{A}$ and the components of $\mathbf{r}$ in the two different coordinate systems are expressed by the column matrices, $\mathbf{Y}$ and $\mathbf{X}$.

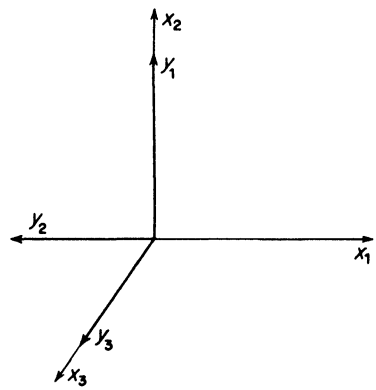


Figure 2.1 Rotation through 90° about the x_3 axis produces a new set of coordinates y_1 whose orientation with respect to the original coordinate system is shown above.

Now we might reasonably regard $\mathbf{A}$ as a rotation operator which generates a new set of coordinate axes y_i by rotating the original set x_i through some angle about some axis. The magnitude of the vector $\mathbf{r}$ remains invariant under the transformation, but its components vary according to Eq. (2.4). For example, if we wished to generate the set y_i from the set x_i by rotation through 90° about the x_3 axis, counterclockwise as viewed from a positive x_3 direction looking toward the origin (see Fig. 2.1), the rotation

matrix would be

$$\begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Successive rotations are accomplished by multiplying by successive matrix rotation operators. In the previous case, a subsequent rotation about the y_1 axis of 90° to produce a coordinate system z_i could be accomplished by the product of the rotation operators

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

A general rotation in space may be accomplished by successive transformations of coordinates

$$\mathbf{Y} = \mathbf{AX}, \qquad \mathbf{Z} = \mathbf{BY}, \qquad \mathbf{X}' = \mathbf{CZ} = \mathbf{CBAX}$$

In Fig. 2.2, these rotations are through an angle θ about the x_2 axis, ϕ about the y_3 axis, and ψ about the z_2 axis. The angles are all measured positive in a counterclockwise direction as viewed from a positive position along the axis of rotation.

The transformation matrices are

$$\mathbf{A} = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \tag{2.5a}$$

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & \sin\phi \\ 0 & -\sin\phi & \cos\phi \end{bmatrix} \tag{2.5b}$$

$$\mathbf{C} = \begin{bmatrix} \cos\psi & \sin\psi & 1 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.5c}$$

If the components of a vector transform under a general rotation of the coordinates according to $\mathbf{Y} = \mathbf{AX}$, it is equally valid to think of the coordinate axes as remaining fixed while the vector suffers an equal but opposite rotation. Thus the matrix equation $\mathbf{Y} = \mathbf{AX}$ can be regarded either as generating a new coordinate system y_i by a rotation of the x_i coordinate system, or as generating a new vector $\mathbf{r}'$. If the components of $\mathbf{r}'$ form the column matrix $\mathbf{Y}$ and the components of $\mathbf{r}$ form the column matrix $\mathbf{X}$ and if $\mathbf{r}'$ is generated by a rotation of $\mathbf{r}$, then the operation $\mathbf{r}' = \mathbf{Ar}$ is equivalent to $\mathbf{Y} = \mathbf{AX}$. Here the rotation produced by $\mathbf{A}$ operating upon $\mathbf{r}$ is equal and opposite to that produced by $\mathbf{A}$ operating upon a set of coordinate axes.

If **A** is a rotation operator, then $\mathbf{A}^{-1}$ will act as a rotation operator which will produce a rotation equal in magnitude but opposite in sense to that

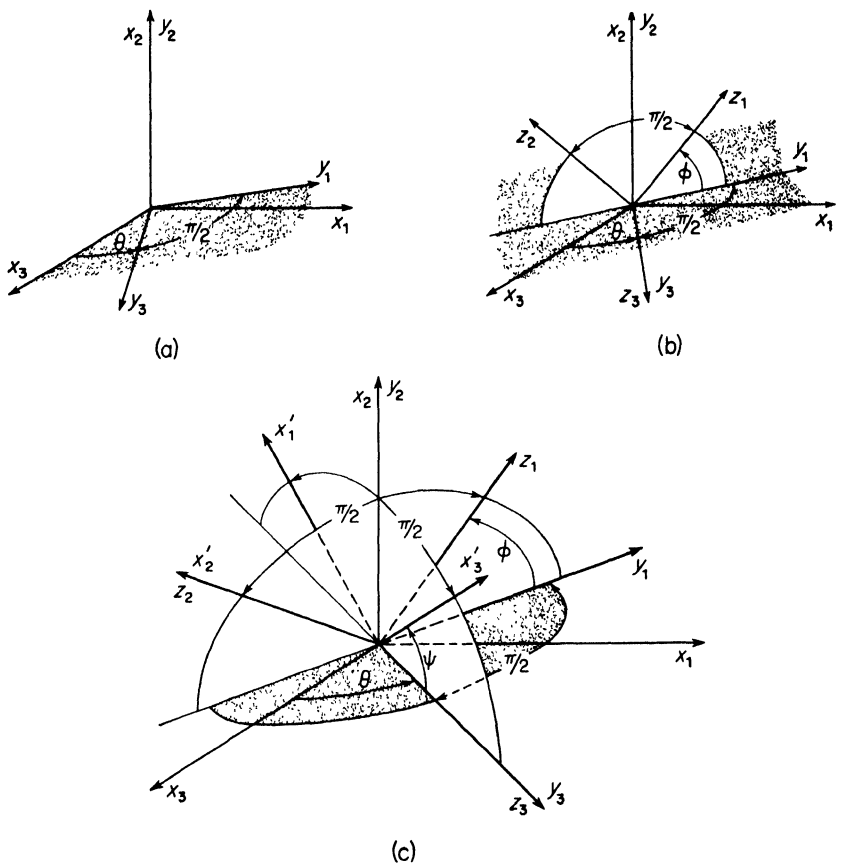


Figure 2.2 The set of matrices appearing in Eqs. 2.5 provide for a general rotation in space where the angles involved in the matrices are illustrated in this figure.

produced by **A**. In this instance, the inverse of **A** is particularly easy to calculate since $\mathbf{A}^{-1}$ equals $\tilde{\mathbf{A}}$.

2.4 The diagonalization of a matrix

Every student of engineering is aware that a general expression for stress at a point expressed in an arbitrary coordinate system contains six components, three of which are shears and three of which are tensions. In that system the

stress components may be represented by a matrix

$$\begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_{zz} \end{bmatrix}$$

Furthermore, there is a particular orientation of the coordinate system in which the shear components disappear. The axes of this coordinate system are called the *principal axes* of stress and the matrix representing the stress referred to the principal axes is

$$\begin{bmatrix} \sigma''_{xx} & 0 & 0 \\ 0 & \sigma''_{yy} & 0 \\ 0 & 0 & \sigma''_{zz} \end{bmatrix}$$

This is called a *diagonal matrix.* Questions now arise about the relationship existing between the elements of the diagonal matrix and those of the original matrix as well as about the orientation of the axes in which the matrix of the stress components is diagonal with respect to the original coordinate axes.

The more intuitive reader will be tempted to try to find some transformation matrix such that a simple rotation of the axes will make the nondiagonal terms zero. Let us recall that when such transformations were applied to vectors a certain reciprocity existed, inasmuch as the same operator could be used to rotate the coordinate system in one sense, or to rotate the vector in the opposite sense. This reciprocity does not exist for quantities such as stress, since the stress components are components of a tensor rather than a vector. It is true that a simple rotation of the axes can be used to reduce the stress matrix to its diagonal form, but this will not suffice to relate the components in the new coordinate system to those in the original system. In order to solve this problem completely, it is necessary to introduce an additional type of transformation called a *similarity transformation* which we develop in the following way.

The vector $\mathbf{r}$, whose components are X_i, can be transformed into another vector $\mathbf{R}$, whose components (in the same coordinate system) are X'_i, by a transformation

$$\mathbf{Ar} = \mathbf{R}$$

If $\mathbf{A}$ is orthogonal,[1] the transformation will consist of a simple rotation of $\mathbf{r}$, but in the more general case, $\mathbf{r}$ may change both its magnitude and direction under this transformation.

Now, if the coordinate system itself is subjected to a transformation, whose matrix is $\mathbf{B}$, so that the components of $\mathbf{r}$ in the new coordinate system are U_i and those of $\mathbf{R}$ are U'_i, what matrix operation will transform U_i into U'_i? The matrices of the components of $\mathbf{R}$ and $\mathbf{r}$ in the original coordinate systems

[1] $\mathbf{A}$ is orthogonal if $a_{ij}a_{jk} = \delta_{ik}$.

are related through

$$(X_i') = \mathbf{A}(X_i)$$

and the transformation to the new coordinate system may be performed by the operation

$$\mathbf{B}(X_i') = \mathbf{BA}(X_i) = (U_i') \tag{2.6}$$

Similarly, the components of $\mathbf{r}$ in the old coordinate system are represented by the matrix (X_i) and the transformation into the new coordinate system gives

$$\mathbf{B}(X_i) = (U_i) \tag{2.7}$$

We may operate upon Eq. (2.7) with the matrix $\mathbf{B}^{-1}$, which is always known to exist, to give

$$(X_i) = \mathbf{B}^{-1}(U_i)$$

and upon Eq. (2.6) with $\mathbf{B}^{-1}$, then with $\mathbf{A}^{-1}$, to give

$$(X_i) = \mathbf{A}^{-1}\mathbf{B}^{-1}(U_i')$$

Hence

$$\mathbf{B}^{-1}(U_i) = \mathbf{A}^{-1}\mathbf{B}^{-1}(U_i')$$

which by operating first with $\mathbf{A}$, then with $\mathbf{B}$, gives

$$\mathbf{BAB}^{-1}(U_i) = (U_i') \tag{2.8}$$

The operation $\mathbf{BAB}^{-1}$ is called a *similarity transformation.* Taken as an operator, $\mathbf{BAB}^{-1}$ will produce the same effect upon a vector $\mathbf{r}$ expressed in terms of its components in the new coordinate system as was produced by $\mathbf{A}$ operating upon $\mathbf{r}$ expressed in terms of its components in the original coordinate system. The new coordinate system is obtained by operating upon the original system with the rotation operator $\mathbf{B}$.

Unfortunately, a development of this type tends to focus all of the attention of the reader upon what the operator $\mathbf{BAB}^{-1}$ is doing to the vector $\mathbf{r}$. Actually, the quantity that is of greatest interest is the matrix $\mathbf{A}$ itself. You will recall that our initial objective was to find a coordinate system in which the shear components of stress are zero. The matrix $\mathbf{A}$ could be the matrix of the components of stress in an arbitrary coordinate system. If we can find some transformation such that

$$\mathbf{BAB}^{-1} = \mathbf{S} \tag{2.9}$$

where $\mathbf{S}$ is a diagonal matrix, we shall have achieved this objective.

In general terms, we seek a matrix

$$\mathbf{S} = \begin{bmatrix} \lambda_1 & 0 & 0 & \dots \\ 0 & \lambda_2 & 0 & \dots \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots \lambda_n \end{bmatrix}$$

such that **Sr** produces the same transformation of **r** when it is represented by its components in the new coordinate system that **Ar** produces when **r** is represented by its components in the original coordinate system.

Letting $\mathbf{C}^{-1} = \mathbf{B}$, Eq. (2.9) becomes

$$\mathbf{C}^{-1}\mathbf{A}\mathbf{C} = \mathbf{S} \tag{2.10}$$

Premultiplying both sides by **C** gives

$$\mathbf{A}\mathbf{C} = \mathbf{C}\mathbf{S}$$

which can be expanded to produce the set of equations

$$a_{ij}c_{jk} = c_{ik}s_{kj} \tag{2.11}$$

$$s_{jk} = 0, \qquad j \neq k \tag{a}$$

$$s_{jk} = \lambda_k, \qquad j = k \tag{b}$$

For a particular value of k, expansion of Eq. (2.11a) results in the set of linear homogeneous equations

$$\begin{aligned}
a_{11}c_{1k} + a_{12}c_{2k} + \ldots + a_{1n}c_{nk} &= c_{1k}\lambda_{(k)} \\
a_{21}c_{1k} + a_{22}c_{2k} + \ldots + a_{2n}c_{nk} &= c_{2k}\lambda_{(k)} \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
a_{n1}c_{1k} + a_{n2}c_{2k} + \ldots + a_{nn}c_{nk} &= c_{nk}\lambda_{(k)}
\end{aligned}$$

If the set of equations is to have any solutions other than the trivial one of all c's equal to zero, the determinant of the coefficients of the c's will be equal to zero. Thus

$$\begin{vmatrix}
a_{11} - \lambda_k & a_{12} & \ldots & a_{1n} \\
a_{21} & a_{22} - \lambda_k & \ldots & a_{2n} \\
\ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\
a_{n1} & a_{n2} & \ldots & a_{nn} - \lambda_k
\end{vmatrix} = 0 \tag{2.12}$$

Expansion of this determinant will give n roots, λ_1, λ_2, etc. Equation (2.12) is called the *characteristic equation* of the matrix **A**, and $\lambda_1, \lambda_2, \ldots, \lambda_n$ are called the *characteristic values* or *eigenvalues* of the matrix **A**.

Having obtained the elements of **S**, we have completed half of our problem. That is, we have a relationship between the elements of the diagonal matrix **S** and the elements of **A**. The remaining part of the problem is to find the orientation of the new coordinate system with respect to the old.

It will be recalled that the c's in Eq. (2.10) are the elements of a matrix **C** which is the inverse of **B**. **B** itself is the rotation operator which was used to rotate the original coordinate system into a position such that the matrix **A** was diagonalized. Consequently, the matrix **C** will provide the inverse operation. If the set y_i represents the set of coordinates in which **A** is diagonal

and if x_i represents the original coordinate system, then

$$(x_i) = \mathbf{C}(y_i)$$

c_{11}, c_{12}, and c_{13} will be proportional to the direction cosines of the y_1 axis with respect to the original coordinate system. Similarly, c_{21}, c_{22}, and c_{23} will be proportional to the direction cosines of the y_2 axis with respect to the x_i system, etc.

Since by this time the elements of (S_{jk}) in Eq. (2.11a) are known from Eq. (2.12) and the elements a_{ij} were known initially, we may attempt to find the elements of the **C** matrix by substitution into Eq. (2.11a). There are difficulties inherent in this procedure. First, even though a set of nine equations and nine unknowns will result, this set is of such a nature that there will be an infinite number of solutions. This can be circumvented by supplying one orthogonality condition such as

$$c_{11}^2 + c_{21}^2 + c_{31}^2 = 1$$

This will permit a solution of the set for all of the elements of the **C** matrix, and c_{11}, c_{21}, and c_{31} will actually be direction cosines as desired. The other elements will, however, only be proportional to sets of direction cosines. These can be normalized by applying more orthogonality conditions. For each additional orthogonality condition that we supply, we may eliminate one equation involving an eigenvalue.

Consequently, if we wish to find a set of solutions for the matrix (C_{ij}), where all of these are direction cosines rather than sets proportional to direction cosines, we may solve the set of equations

$$\begin{aligned}
(a_{11} - \lambda_1)c_{11} + a_{12}c_{21} + a_{13}c_{31} &= 0 \\
a_{21}c_{11} + (a_{22} - \lambda_1)c_{21} + a_{33}c_{31} &= 0 \\
c_{11}^2 + c_{21}^2 + c_{31}^2 &= 1 \\
(a_{11} - \lambda_2)c_{12} + a_{12}c_{22} + a_{32}c_{32} &= 0 \\
a_{11}c_{12} + (a_{12} - \lambda_2)c_{22} + a_{13}c_{32} &= 0 \\
c_{12}^2 + c_{22}^2 + c_{32}^2 &= 1 \\
(a_{11} - \lambda_3)c_{13} + a_{12}c_{23} + a_{13}c_{33} &= 0 \\
a_{11}c_{13} + (a_{12} - \lambda_3)c_{23} + a_{13}c_{33} &= 0 \\
c_{13}^2 + c_{23}^2 + c_{33}^2 &= 1
\end{aligned}$$

If the eigenvalues are not all distinct, then two of the y_i coordinates may be located anywhere on a plane perpendicular to the third axis. The direction cosines of the third coordinate will then suffice to provide a coordinate system in which the matrix is diagonal.

Example 2.2. It is desired to diagonalize the following matrix and to find the orientation of the principal axes with respect to the original coordinate system:

$$\begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

The characteristic equation is obtained from the equation

$$\begin{vmatrix} 1-\lambda & 2 & 0 \\ 2 & 1-\lambda & 1 \\ 0 & 1 & 1-\lambda \end{vmatrix} = 0$$

which reduces to $-\lambda^3 + 3\lambda^2 + 2\lambda - 4 = 0$ with roots $\lambda = 1, 1 \pm \sqrt{5}$. The diagonal form of the matrix is

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 3.21 & 0 \\ 0 & 0 & -1.23 \end{bmatrix}$$

We may compute c_{11}, c_{21}, and c_{31} from

$$(1-1)c_{11} + 2c_{21} + c_{31} = 0 \qquad \text{(a)}$$

$$2c_{11} + (2-1)c_{21} + c_{31} = 0 \qquad \text{(b)}$$

$$c_{11}^2 + c_{21}^2 + c_{31}^2 = 1 \qquad \text{(c)}$$

From Eqs. (a), (b), and (c) we get the following values: $c_{21} = 0$, $c_{11} = 1/\sqrt{5}$, and $c_{31} = -2/\sqrt{5}$. Except that the arithmetic becomes harder, we could also have computed the other direction cosines in a similar manner.

PROBLEMS

2.1 Perform the indicated operations:

(a) $\begin{bmatrix} 2 & 1 \\ -1 & 3 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 4 & 2 \end{bmatrix} =$

(b) $\begin{bmatrix} 3 & 1 & 1 \\ -1 & 2 & 2 \\ 1 & -2 & 2 \end{bmatrix}\begin{bmatrix} 1 & 4 & 1 \\ 4 & 1 & 4 \\ 1 & 4 & 1 \end{bmatrix} =$

(c) $[1 \quad 3 \quad 2]\begin{bmatrix} 4 \\ -1 \\ 2 \end{bmatrix} =$

(d) $\begin{bmatrix} 2 & 1 \\ 3 & 5 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} =$

(e) $\begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix}\begin{bmatrix} 3 & 1 \\ 4 & 2 \end{bmatrix}\begin{bmatrix} 2 & 3 \\ 3 & 1 \end{bmatrix} =$

2.2 Compute the inverse of the following matrices. *Note:* recall that the inverse of an orthogonal matrix is equal to its transpose. It may save time to test for orthogonality before each computation.

(a) $\begin{bmatrix} 2 & 1 & 0 \\ 3 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

(b) $\begin{bmatrix} \sqrt{\frac{3}{2}} & \frac{1}{2} & 0 \\ -\frac{1}{2} & \sqrt{\frac{3}{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$

(c) $\begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}$

(d) $\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$

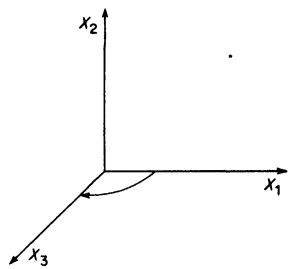
2.3 Prove that the matrix $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ remains unaltered by an arbitrary similarity transformation.

2.4 Diagonalize the matrix

$$\begin{bmatrix} 1 & 0 & -3 \\ 0 & 1 & 0 \\ -3 & 0 & -10 \end{bmatrix}$$

2.5 Write the transformation matrices corresponding to the following operations performed upon a set of Cartesian coordinates:

(a) Rotation through 90° about the x_2 axis in the sense given by



(b) Rotation through 180° about the x_3 axis.

(c) Reflection through the x_1x_3 plane.

(d) Reflection through the origin.

(e) Reflection through the x_2x_3 plane followed by rotation about the x_1 axis through an angle of 60° in the sense indicated by the diagram of Prob. 2.5a.

2.6 Let a_{ij} be the elements of the matrix **A**. Derive a set of orthogonality relationships between the elements a_{ij} from the condition that $\mathbf{A} = \mathbf{A}^{-1}$.

3

INTRODUCTION TO TENSOR ANALYSIS

3.1 Introduction

Certain material properties are by nature tensor quantities. Quantitative calculations based upon these properties require that they be treated as tensors rather than as scalars, and a complete insight into their nature requires an understanding of the concept of a tensor. The reader has almost certainly encountered tensor quantities previously although they may not have been labeled as such. Before presenting a formal mathematical definition, it may be helpful to examine the elements of a fairly familiar physical quantity which is a tensor. Figure 3.1 shows an elementary cube which has certain forces acting upon its faces. These forces are shown on only three of the faces, but because each face has equal and opposite forces acting upon the opposite face, the cube is not subjected to a translational acceleration. If we further stipulate that $F_{ij} = F_{ji}$, etc., then the sum of the moments about any point will be zero and the cube will be in a state of static equilibrium. Consequently, the only effect this force system can have upon the cube is to distort it.

Now, for any given material, distortion will depend upon the magnitude of each force, the area of the face it acts upon, and whether the force is parallel to or perpendicular to the face. Consequently, if we wish to characterize the action of these forces upon the cube, we must cope with a quantity encompassing the magnitudes and directions of the forces as well as the magnitudes and orientation of the faces. This quantity is called a *stress*. For each pair of opposing faces of the cube there will be three equations which

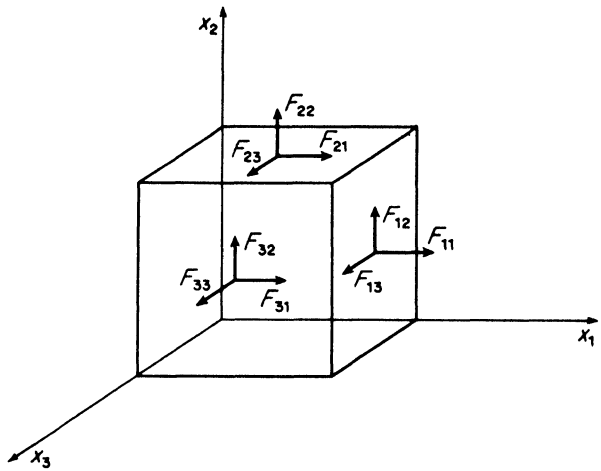


Figure 3.1 Illustration of the surface forces acting upon an elementary cube in a continuous medium.

will characterize the stress acting upon the faces. Thus, for the faces perpendicular to the x_1 axis, we have

$$\sigma_{11}\,dA_1 = F_{11}$$
$$\sigma_{12}\,dA_1 = F_{12}$$
$$\sigma_{13}\,dA_1 = F_{13}$$

The quantity σ_{11} is called a *tensile component* of the stress and tends to produce an elongation in the x_1 direction and contractions in the x_2 and x_3 directions. σ_{12} and σ_{13} are called *shear components;* they tend to displace opposite faces of the cube in parallel directions. Similarly for the other faces, we may write

$$\sigma_{22}\,dA_2 = F_{22}, \qquad \sigma_{33}\,dA_3 = F_{33}$$
$$\sigma_{21}\,dA_2 = F_{21}, \qquad \sigma_{31}\,dA_3 = F_{31}$$
$$\sigma_{23}\,dA_2 = F_{23}, \qquad \sigma_{32}\,dA_3 = F_{32}$$

The stress, as a whole, may be considered an entity having a total of nine components, each of which is involved in some way with a pair of mutually perpendicular directions or is involved in a dual manner with one direction. These nine components are analogous to the components of a vector.

This physical approach to tensors is necessarily rather limited; it has been introduced only for conceptual purposes. Now it is necessary to adopt a more formal procedure. First, we shall deal only with Cartesian tensors in this text, and the definitions that follow must be accepted with the reservation that they are not all applicable to all tensors.

3.2 Transformation properties of tensors

Tensors are classified according to rank; a *tensor* of *rank n* is a quantity which has 3^n components in a three-dimensional Euclidean space which vary in a particular way under a transformation of coordinates. It turns out that this is simply a generalization of the way that the components of a vector may vary under such a transformation. Indeed, a vector is a tensor of rank one. We recall that the relationship between the components of a vector in one coordinate system and those in another coordinate system is obtained through the operation $\mathbf{r}' = \mathbf{Ar}$ where $\mathbf{A}$ is a transformation matrix whose elements are the direction cosines between the new axes and the old. We label the new axes, y_1, y_2, y_3, and the old, x_1, x_2, x_3. Then the transformation matrix is

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

where a_{11} is the cosine of the angle between the axes x_1 and y_1, a_{12} equals the cosine of the angle between the x_2 and y_1 axes, etc. Then the components are

$$r_1' = a_{11}r_1 + a_{12}r_2 + a_{13}r_3$$
$$r_2' = a_{21}r_1 + a_{22}r_2 + a_{23}r_3$$

and so on. For any component we may write

$$r_i' = a_{ij}r_j$$

using the summation convention.

With this background, we may now define a *tensor* of *rank n* as a quantity which transforms under a rotation of the coordinates according to

$$T'_{mno\ldots} = a_{mi}a_{nj}a_{ok}\ldots T_{ijk\ldots} \tag{3.1}$$

where the summation is over all the combinations of values of i, j, k, etc.

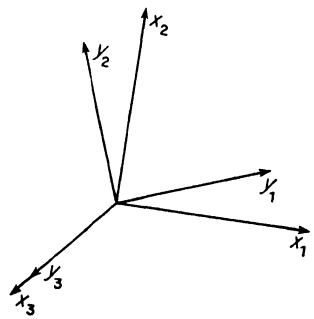
When a tensor is expressed in terms of its components, its rank is immediately apparent from the number of free subscripts used on the components. However, we may often find it convenient to designate a tensor by a boldface capital letter such as $\mathbf{T}$ just as we have designated vectors. In cases where this is convenient, the rank of the tensor either must be known from context or must be stated explicitly. There are many circumstances in which we may need only indicate an operation involving tensors, and in these cases this notation is most convenient.

It is fairly important to realize that the tensor itself is an entity which has an existence independent of the coordinate system in which its components are represented. Its components change under a rotation of the coordinates according to a prescribed formula, but the tensor itself does not change under

such a transformation. Despite this independent existence we can give a tensor a precise mathematical definition only in terms of the number of components it has and the way that these transform under a rotation of the coordinate axes. The nonmathematician seeking a concept of what a tensor "is" is usually not content with the formal definition given here. However, we must distinguish between tensors *per se* and physical quantities that behave in such a way that a complete description of their properties requires that they be represented as tensors. We shall encounter numerous examples of these in later chapters. For the present it is best to live, perhaps uncomfortably, with an abstract definition of a tensor, and to master the intricacies of tensor operations without attempting to attach a physical meaning to them.

In its most general form, the definition as embodied in Eq. (3.1) is fairly hard to understand. Its essential points can be illustrated by considering specific tensor quantities and specific coordinate transformations. The three examples that follow are worked out in detail and should help the reader develop an understanding so that the defining equation has more meaning.

Example 3.1. The figure shows two coordinate systems x_i and y_i having a particular orientation with respect to one another. The components of the vector



A (a tensor of rank one) in the x_i coordinate system are A_1, A_2, and A_3; A_1 being the component in the x_1 direction, etc. We wish to find the components of **A** in the y_i coordinate system. According to Eq. (3.1), the transformation permitting us to compute an arbitrary component A'_m is

$$A'_m = a_{m1}A_1 + a_{m2}A_2 + a_{m3}A_3$$

where a_{m1} is the direction cosine between the x_1 axis and the y_m axis, etc. It is helpful to write the entire transformation matrix

$$(a_{m1}) = \begin{bmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} & 0 \\ +\frac{1}{2} & \frac{\sqrt{3}}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

from which we compute the components of A'_m of **A** in the y_i coordinate system to be

$$A'_1 = \frac{\sqrt{3}}{2} A_1 - \tfrac{1}{2} A_2$$

$$A'_2 = \frac{\sqrt{3}}{2} A_1 + \tfrac{1}{2} A_2$$

$$A'_3 = A_3$$

Example 3.2. Let **T** be a tensor of rank two whose components in the x_i coordinate system are

$$T_{11} = A, \qquad T_{22} = B, \qquad T_{33} = C; \qquad \text{all other components} = 0$$

Transforming to the y_i coordinate system of Example 3.1 we have, from the general transformation formula,

$$T'_{ij} = a_{ik} a_{jl} T_{kl}$$

The result is

$$T'_{11} = \tfrac{3}{4}A + \tfrac{1}{4}B \qquad T'_{12} = +\frac{\sqrt{3}}{4} A - \frac{\sqrt{3}}{4} B \qquad T'_{13} = 0$$

$$T'_{21} = +\frac{\sqrt{3}}{4} A - \frac{\sqrt{3}}{4} B \qquad T'_{22} = \tfrac{1}{4}A + \tfrac{3}{4}B \qquad T'_{23} = 0$$

$$T'_{31} = 0 \qquad T'_{32} = 0 \qquad T'_{33} = C$$

Example 3.3. Let **T** be a tensor of rank three whose components in the x^i coordinate system of Example 3.1 are

$$T_{111} = A, \quad T_{132} = B, \quad T_{133} = 0, \quad T_{121} = B, \quad T_{122} = 0, \quad T_{123} = 0$$
$$T_{131} = 0, \quad T_{112} = 0, \quad T_{113} = 0, \quad T_{211} = 0, \quad T_{212} = C, \quad T_{213} = G$$
$$T_{221} = C, \quad T_{222} = D, \quad T_{223} = 0, \quad T_{231} = G \quad T_{232} = 0, \quad T_{233} = 0$$
$$T_{311} = 0, \quad T_{312} = 0, \quad T_{313} = 0, \quad T_{321} = 0, \quad T_{322} = 0, \quad T_{323} = E$$
$$T_{331} = 0, \quad T_{332} = E, \quad T_{333} = F$$

Let us compute the component T'_{323} in the y_i coordinates of Example 3.1. For a tensor of rank three, the transformation formula is

$$T'_{mno} = a_{mi} a_{nj} a_{ok} T_{ijk}$$

Omitting the zero terms from the summation, we have

$$\begin{aligned} T'_{323} = {} & a_{31}a_{21}a_{31}A + a_{31}a_{23}a_{32}B + a_{31}a_{22}a_{31}B + a_{32}a_{21}a_{32}C \\ & + a_{32}a_{22}a_{31}C + a_{32}a_{23}a_{31}G + a_{32}a_{21}a_{33}G + a_{33}a_{22}a_{33}E \\ & + a_{33}a_{23}a_{33}F + a_{32}a_{22}a_{32}D + a_{33}a_{23}a_{32}E \end{aligned}$$

Substitution from the transformation matrix of Example 3.1 reduces most of these to zero, leaving

$$T'_{323} = \frac{\sqrt{3}}{2} E$$

3.3 Tensor symmetry and tensor addition

If **T** is a tensor of rank two and its components are such that

$$T_{ij} = T_{ji} \tag{3.2}$$

then **T** is called a *symmetric tensor*. Most tensor properties of materials have this characteristic, and the fact that the symmetry (of such a tensor) is not destroyed by a rotation of coordinates will be very useful later in making certain deductions about material properties.

There are also tensors whose components are related by

$$T_{ij} = -T_{ji} \tag{3.3}$$

Those components with equal subscripts will necessarily be zero under these circumstances. Such tensors are variously described as *antisymmetric* or *skew-symmetric*. We shall prove shortly that any tensor may be broken into the sum of a symmetric tensor and an antisymmetric tensor. But first we must define the operation of addition.

The addition of tensors may be encountered in the solution of engineering problems. In particular, this operation occurs in the development of important theorems. Two tensors are added by simply adding their corresponding components. Thus if

$$\mathbf{B} = \mathbf{P} + \mathbf{T}$$

then

$$B_{\ldots ij} = P_{\ldots ij} + T_{\ldots ij} \tag{3.4}$$

The operation is defined only for tensors of equal rank whose components have corresponding subscripts. It could be extended to include three or more tensors, and the operation is commutative and associative. Our only application of this operation is in explaining the following theorem:

> **Theorem I.** Any tensor of rank two may be expressed as the sum of a symmetric and an antisymmetric tensor.

We illustrate this by separating the components T_{ij} of a rank two tensor into a sum and difference:

$$\begin{aligned} T_{ij} &= \frac{T_{ij} + T_{ji}}{2} + \frac{T_{ij} - T_{ji}}{2} = S_{ij} + A_{ij} \\ T_{ji} &= \frac{T_{ji} + T_{ij}}{2} + \frac{T_{ji} - T_{ij}}{2} = S_{ji} + A_{ji} \end{aligned} \tag{3.5}$$

from which we see that

$$S_{ij} = S_{ji} \qquad \text{and} \qquad A_{ij} = -A_{ji}$$

leading to the conclusion that S_{ij} are components of a symmetric tensor and A_{ij} are those of an antisymmetric tensor.

The following theorem, which is presented without proof, represents a very important fact about symmetric tensors of rank two:

> **Theorem II.** There is some orientation of the coordinate axes such that the components of a symmetric tensor of rank two T_{ij} will all be zero for j not equal to i.

The axes of a coordinate system in which this is true are called the *principal axes* of the tensor. One of the many consequences of this theorem is that any symmetric tensor of rank two may be characterized by a maximum of three quantities: namely, its components along the principal axes of the tensor. We may express the components of the tensor in any other coordinate system as a function of these three quantities and the direction cosines between the axes of the old and new coordinates.

Conversely, if we know the components of a symmetric tensor in an arbitrary coordinate system, we may find the components in a set of principal axes by evaluating the λ's in the determinant

$$\begin{vmatrix} T_{11} - \lambda & T_{12} & T_{13} \\ T_{21} & T_{22} - \lambda & T_{23} \\ T_{31} & T_{32} & T_{33} - \lambda \end{vmatrix} = 0 \tag{3.6}$$

This is analogous to the diagonalization of a matrix that was presented in Chap. 2. For many purposes, we may regard a rank two tensor and the matrix of its components as being formally identical. The orientation of the principal axes with respect to the arbitrary coordinates chosen earlier can also be found.

Examples of transformations from a set of principal axes to some other coordinate system and the reverse operation follow.

> **Example.3.4.** The components of strain at a point in a structural member in a principal coordinate system are
>
> $$T_{11} = 2 \times 10^{-6}, \qquad T_{22} = 0, \qquad T_{33} = 5 \times 10^{-6}$$
>
> Labeling this coordinate system x_i, find the components of the strain in a coordinate system y_i in which the x_3 and y_3 axes coincide and the angle between the x_1 and the y_1 axis is 45°. The transformation matrix is
>
> $$\begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Since the coordinates x_i are principal axes, T_{12}, T_{13}, and T_{23} are zero and the transformation equation reduces to

$$T'_{mn} = a_{m1}a_{n1}T_{11} + a_{m3}a_{n3}T_{33}$$

Hence

$$T'_{12} = a_{11}a_{21}T_{11} + a_{13}a_{23}T_{33} = +10^{-6}$$

$$T'_{13} = a_{11}a_{31}T_{11} + a_{13}a_{33}T_{33} = 0$$

$$T'_{32} = a_{31}a_{21}T_{11} + a_{33}a_{23}T_{33} = 0$$

$$T'_{11} = a_{11}^2T_{11} + a_{13}^2T_{33} = \frac{T_{11}}{2} = 10^{-6}$$

$$T'_{22} = a_{21}^2T_{11} + a_{23}^2T_{33} = \frac{T_{11}}{2} = 10^{-6}$$

$$T'_{33} = a_{31}^2T_{11} + a_{33}^2T_{33} = T_{33} = 5 \times 10^{-6}$$

The strain tensor is symmetric, consequently

$$T'_{21} = T'_{12}, \qquad T'_{13} = T'_{31}, \qquad T'_{23} = T'_{32}$$

Example 3.5. The components of a symmetric tensor of rank two in some coordinate system are $T_{12} = T_{21} = 3$, $T_{13} = T_{31} = 5$,

$$T_{11} = 2, \quad T_{22} = 4, \quad T_{23} = T_{32} = 6, \quad T_{33} = 8$$

Find the components of the tensor in a set of principal axes.

This is accomplished by the evaluation of the determinant equation

$$\begin{vmatrix} 2-\lambda & 3 & 5 \\ 3 & 4-\lambda & 6 \\ 5 & 6 & 8-\lambda \end{vmatrix} = 0$$

which gives

$$(2-\lambda)(32 - 12\lambda + \lambda^2 - 36) - 3[3(8-\lambda) - 30] + 5(18 - 20 + 5\lambda) = 0$$

Expanding and rearranging terms gives

$$-\lambda^3 + 14\lambda^2 + 14\lambda = 0$$

whose roots are 0, -0.92, and 14.92.

3.4 The contraction operation

In Sec. 1.3 an operation called the dot product was defined for vectors by the equation

$$\mathbf{A} \cdot \mathbf{B} = a_ib_i \tag{3.7}$$

(**A** and **B** are rank one tensors.) We may readily extend this operation to include tensors. If **T** is a tensor of any rank and **A** is a tensor of any rank, then

$$\mathbf{T} \cdot \mathbf{A} = T_{\ldots jkl}A_{lmn\ldots} \tag{3.8}$$

To cite explicit examples, let **T** be a rank two tensor and **A** be a vector, then

$$\mathbf{T} \cdot \mathbf{A} = T_{ij}A_j$$

The quantities $T_{ij}A_j$ are components of a vector. Written in extended form, these components are

$$V_1 = T_{11}A_1 + T_{12}A_2 + T_{13}A_3$$
$$V_2 = T_{21}A_1 + T_{22}A_2 + T_{23}A_3$$
$$V_3 = T_{31}A_1 + T_{32}A_2 + T_{33}A_3$$

If both **T** and **A** are rank two tensors, then the product of the operation $\mathbf{T} \cdot \mathbf{A}$ is also a rank two tensor whose components are

$$Q_{ik} = T_{ij}A_{jk}$$

We may also form a double contraction that we symbolize by $\mathbf{T} : \mathbf{A}$. Both **T** and **A** must have a rank of two or more, and the result of a double contraction will be a tensor of a rank equal to the sum of the ranks of **T** and **A** minus four. The operation is defined by the statement that if

$$\mathbf{B} = \mathbf{T} : \mathbf{A}$$

then

$$B_{\ldots hk\ldots} = T_{\ldots hij}A_{ijk\ldots} \tag{3.9}$$

If both **T** and **A** are rank two tensors, then $\mathbf{T} : \mathbf{A}$ is a scalar equal to

$$T_{ij}A_{ij} = T_{11}A_{11} + T_{12}A_{12} + T_{13}A_{13} + T_{21}A_{21} + T_{22}A_{22} + T_{23}A_{23} + T_{31}A_{31} + T_{32}A_{32} + T_{33}A_{33}$$

If **T** is a rank three tensor and **A** a rank two tensor, then $\mathbf{T} : \mathbf{A}$ is a vector whose components are $T_{ijk}A_{jk}$.

Occasionally it is convenient to write an equation in the symbolic form

$$\mathbf{A} = \frac{\mathbf{B}}{\mathbf{C}}$$

where **A**, **B**, and **C** are tensors of any rank so long as the rank of **A** plus the rank of **C** is equal to or greater than the rank of **B**. This symbolic equation means that **B** is equal to a contraction between **A** and **C**, the degree of contraction being governed by the rank of the tensors involved. Thus the equation could represent any of the following operations:

$\mathbf{AC} = \mathbf{B}$	when the rank of **A** plus the rank of **C** equals the rank of **B**
$\mathbf{A} \cdot \mathbf{C} = \mathbf{B}$	when the rank of **A** plus the rank of **C** equals the rank of **B** plus one
$\mathbf{A} : \mathbf{C} = \mathbf{B}$	when the rank of **A** plus the rank of **C** equals the rank of **B** plus two.

3.5 The dyadic product

The contraction operation described previously usually (though not necessarily) produces a tensor of rank two lower than that of the sum of the ranks of the participants in the operation. We may also define a product called a *dyadic product* which produces tensors of higher rank. Let **A** and **B** be tensors of any rank. The dyadic product **AB** produces a tensor whose rank is equal to the sum of the ranks of **A** and **B** and whose components are

$$A_{ijk\ldots}B_{opq\ldots}$$

For example, if **A** and **B** are both vectors, then their dyadic product is a rank two tensor whose components are

$$\begin{array}{lll} A_1B_1, & A_1B_2, & A_1B_3 \\ A_2B_1, & A_2B_2, & A_2B_3 \\ A_3B_1, & A_3B_2, & A_3B_3 \end{array}$$

Note that $\mathbf{AB} \neq \mathbf{BA}$.

We may combine the dyadic product with a contraction by operations such as

$$\mathbf{C} \cdot \mathbf{AB} = (\mathbf{C} \cdot \mathbf{A})\mathbf{B}$$

and we may also use the vector operator in combination with tensors to form dyadic products. We shall find later that when a continuous medium is deformed, the dyad $\nabla\mathbf{D}$ where **D** is a properly defined vector will help describe the deformation of the medium.

3.6 Differentiation

Let us suppose that the set of 3^n quantities $A_{\ldots ij\ldots}$ are the components of a tensor of rank n. Let us further suppose that each of these is a function of 3^m independent variables $B_{\ldots kl\ldots}$ which are components of a tensor of rank m. We may form a set of 3^{m+n} partial derivatives

$$\frac{\partial A_{\ldots ij\ldots}}{\partial B_{\ldots kl\ldots}}\Bigg)_{\substack{\text{All other components} \\ \text{of } \mathbf{B} \text{ held constant}}}$$

which are components of a tensor of rank $m + n$.

Thus, for example, if ϕ is a tensor of rank zero (i.e., a scalar) and **A** is a tensor of rank one, and if A_1, A_2, and A_3 are all functions of ϕ, then the set $\partial A_1/\partial\phi\big)_{A_2,A_3}$, $\partial A_2/\partial\phi\big)_{A_1,A_3}$, and $\partial A_3/\partial\phi\big)_{A_1,A_2}$, are components of a tensor of rank one. Conversely, if ϕ is a function of the independent variables A_1, A_2, A_3 and these are components of a tensor of rank one, then the set $\partial\phi/\partial A_i\big)_{A_j,A_k}$ are also components of a tensor of rank one.

By way of further illustration, let **A** be a tensor of rank three and **B** a tensor of rank two. If each component of **A** is a function of all of the variables $B_{11}, B_{12}, \ldots, B_{23}$, then the 243 partial derivatives

$$\frac{\partial A_{ijk}}{\partial B_{lm}}\Bigg)_{\substack{\text{All other components}\\ \text{of } \mathbf{B} \text{ held constant}}}$$

are components of a tensor of rank five.

Although it is improper to speak of a derivative of one tensor with respect to another, it is notationally convenient to define the expression

$$\frac{d\mathbf{A}}{d\mathbf{B}} \tag{3.10}$$

as meaning a tensor **C** whose rank is equal to the sum of the ranks of **A** and **B**, and whose components are

$$C_{\ldots ijkl\ldots} = \frac{\partial A_{\ldots ij}}{\partial B_{kl\ldots}}\Bigg)_{\substack{\text{All other } \mathbf{B}\text{'s}\\ \text{held constant}}} \tag{3.11}$$

If the components of **A** are functions of two or more different sets of variables, each set being components of other tensors, then the symbol $\partial\mathbf{A}/\partial\mathbf{B}$ will have an analogous meaning. For example, let **A**, **B**, and **C** be tensors of rank two and let

$$A_{ij} = A_{ij}(B_{11}, B_{22}, \ldots, C_{11}, C_{12}, \ldots)$$

then $\partial\mathbf{A}/\partial\mathbf{B}\Big)_{\mathbf{C}}$ is a tensor **D** of rank four whose components are

$$D_{ijkl} = \frac{\partial A_{ij}}{\partial B_{kl}}\Bigg)_{\substack{\text{All other components of } \mathbf{B} \text{ and}\\ \text{all components of } \mathbf{C} = \text{const.}}}$$

3.7 Unit tensors

There are two rather unique tensors which occur repeatedly in analysis and which have the rather unusual feature of having their *components* remain invariant under an orthogonal transformation. We have already encountered these in Chap. 1, but did not treat them as tensors. The *Kronecker* δ is a rank two tensor whose components are defined by

$$\begin{aligned} \delta_{ij} &= 1, \qquad i = j \\ &= 0 \qquad i \neq j \end{aligned} \tag{3.12}$$

In addition to the applications already encountered, the Kronecker δ provides a means of defining the inverse of a rank two tensor. If **T** is a rank two tensor, then we may define another rank two tensor **Q** by the equation

$$\mathbf{T} \cdot \mathbf{Q} = \boldsymbol{\delta}$$

where **Q** is the inverse of **T**.

Upon expansion it will be found that the components of **T** and **Q** bear the same relationship to one another that the elements of a matrix and the inverse of the same matrix bear to one another. The operation $\mathbf{T} \cdot \mathbf{Q}$ is commutative in contrast with the usual state of affairs.

The primary utility of having defined an inverse is in solving an equation such as

$$C_{ij}D_j = E_{ik}F_k$$

for either D_j or F_k. If **B** is the inverse of **C**, then

$$B_{oi}C_{ij}D_j = \delta_{oj}D_j = B_{oi}E_{ik}F_k$$

which is the desired result.

The other tensor whose components remain invariant under an orthogonal transformation is the *Levi-Civita density*. This is a rank three tensor whose components are defined by u_{ijk} equal 1 when i, j, k are ordered in the sequence 12312... and equal -1 when i, j, k are ordered in the sequence 32132... and equal 0 when any pair or all subscripts are equal.

We have already used the Levi-Civita density in the expansion of a vector cross product $\mathbf{V} = \mathbf{A} \times \mathbf{B}$ into its components in the form

$$V_i = u_{ijk}A_jB_k$$

3.8 Symmetry of tensors of rank higher than two

The components of tensors of rank two are distinguished by two subscripts. There are only two possible ways of arranging these subscripts and a symmetric tensor can be readily and unambiguously defined. But when we consider tensors of higher rank, the matter is no longer so simple. Should we require that $T_{ijk} = T_{ikj}$ as our criterion of symmetry? Or should it be $T_{kji} = T_{ijk}$ or $T_{ijk} = T_{jik}$? Let us attempt an answer by approaching the problem indirectly.

First, let us suppose that **T** is a tensor of rank three, and we know, perhaps from physical reasoning, that **T** represents a quantity such that $\mathbf{T} \cdot \mathbf{A}$ where **A** is a vector produces a symmetric tensor **P**. If this is the case, then

$$P_{ij} = T_{ijk}A_k = P_{ji} = T_{jik}A_k$$

Therefore we would have

$$T_{ijk} = T_{jik}$$

This would reduce the number of independent components of **T** from 27 to 18 and is a possible criterion for the symmetry of **T**. However, suppose that instead of $\mathbf{P} = \mathbf{T} \cdot \mathbf{A}$, the relationship were $\mathbf{P} = \mathbf{A} \cdot \mathbf{T}$. Then

$$P_{ij} = T_{kij}A_k = P_{ji} = T_{kji}A_k$$

and we would have

$$T_{kij} = T_{kji}$$

Actually either criterion could serve as a definition of symmetry and so could the relationship $T_{ijk} = T_{kji}$ for that matter. In this sense a unique definition of symmetry is not available for tensors of rank greater than two. Perhaps, this in itself is not too important. The major point is that when a tensor of rank higher than two is subjected to an operation resulting in a symmetric tensor of rank two, certain components of the original tensor will necessarily be equal.

To carry this thought further, if $\mathbf{P} = \mathbf{A} \cdot \mathbf{T}$ and if $\mathbf{P}$ is symmetric, then there is some coordinate system for which $\mathbf{P}_{ij} = 0$, $i \neq j$, when $\mathbf{A}$ is arbitrary. In this coordinate system

$$A_k T_{kij} = 0 \qquad j \neq i$$

and $\mathbf{T}$ would have only nine nonzero components. Consequently if a tensor of rank three has the properties ascribed to $\mathbf{T}$, that is, a symmetric tensor is derivable from a single contraction of $\mathbf{T}$, then all 27 components of $\mathbf{T}$ in an arbitrary coordinate system may be expressed as functions of a maximum of nine independent quantities.

Tensors of rank four are encountered quite frequently in working with materials. If a contraction of such a tensor produces a symmetric tensor of rank two, we may again explore the restrictions that will be imposed upon its components in a manner similar to that used for tensors of rank three. Let $\mathbf{T}$ be a tensor of rank four, and let $\mathbf{S}$ be a symmetric tensor of rank two such that

$$S_{ij} = T_{ijkm} P_{km}$$

If S_{ij} is symmetric, then for arbitrary P_{km}

$$S_{ij} = T_{ijkm} P_{km} = S_{ji} = T_{jikm} P_{km}$$

Therefore

$$T_{ijkm} = T_{jikm}$$

If P_{km} is also symmetric, then

$$T_{ijkm} = T_{jikm} = T_{ijmk} = T_{jimk}$$

This reduces the total number of independent components of $\mathbf{T}$ from 81 to 36.

In the case of elastic compliances or elastic rigidities, we shall find later that there is one further restriction that arises from physical reasoning. That is,

$$T_{ijkm} = T_{kmij}$$

which reduces the number of distinct components still further to 21. These in turn may each be expressed as functions of 18 independent quantities, since a coordinate system may be chosen so that

$$T_{ijkm} = 0, \qquad j \neq k$$

3.9 Summary

A tensor is defined only in terms of the number of its components and the way that these transform under a rotation of the coordinate axes. Many engineers and physicists will find this definition a bit too abstract, but it will assume a more concrete significance after one has had some practice in working with physical quantities which are described in mathematical terms as tensors. All of the operations that we have defined in this chapter will be used repeatedly in the arguments presented in later chapters and should be mastered rather thoroughly. The concept of tensor symmetry also will be used extensively, and important deductions about properties of materials will depend upon the fact that there are certain preferred orientations of coordinates in which tensors of rank two have only three nonzero components, and in which the number of nonzero components of tensors of rank three and four are, respectively, nine and eighteen.

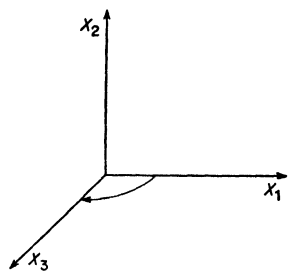
PROBLEMS

3.1 When referred to a specified set of Cartesian coordinate axes, the components of a rank two tensor **T** are

$$T_{11} = 1 \quad T_{12} = -2 \quad T_{13} = 3 \quad T_{21} = 4$$
$$T_{23} = 2 \quad T_{22} = 0 \quad T_{31} = 4 \quad T_{32} = 2 \quad T_{33} = 3$$

(a) Find an anti-symmetric tensor **A** and a symmetric tensor **S** such that $\mathbf{T} = \mathbf{A} + \mathbf{S}$.

(b) Evaluate the components of **T** in coordinate systems derived from original by (1) rotation through 180° about the x_1 axis, (2) reflection through the x_2x_3 plane, and (3) rotation through 90° about the x_2 axis in the sense



(c) Let **A** be a rank one tensor whose components in the original coordinate system are $A_1 = 3$, $A_2 = -2$, $A_3 = 0$. Evaluate the scalar $\mathbf{A} \cdot \mathbf{T} \cdot \mathbf{A}$ and the components of the vectors $\mathbf{A} \cdot \mathbf{T}$ and $\mathbf{T} \cdot \mathbf{A}$.

(d) Let **B** be a rank two tensor whose components are

$$B_{11} = 2 \quad B_{22} = 4 \quad B_{33} = 6$$
$$B_{12} = B_{21} = B_{13} = B_{31} = B_{23} = B_{32} = 0$$

Evaluate the scalars $\mathbf{T}:\mathbf{B}$ and $\mathbf{B}:\mathbf{T}$ and the components of the rank two tensors $\mathbf{T}\cdot\mathbf{B}$ and $\mathbf{B}\cdot\mathbf{T}$.

(e) Let $\mathbf{A}$ be the vector of part c, and $\mathbf{C}$ be a vector whose components are $C_1 = 3$, $C_2 = 1$, $C_3 = 1$. Evaluate the components of $\mathbf{T}\cdot(\mathbf{A}\times\mathbf{C})$.

(f) Evaluate the components of the dyad $\mathbf{AC}$.

3.2 The stress at a point in a material has the following components when referred to a particular coordinate system:

$$\sigma_{11} = 10^3 \text{ psi} \qquad \sigma_{12} = \sigma_{21} = 2\times 10^3 \text{ psi} \qquad \sigma_{22} = 0$$
$$\sigma_{32} = \sigma_{23} = 0 \qquad \sigma_{31} = \sigma_{13} = 0.5\times 10^3 \text{ psi} \qquad \sigma_{33} = 0$$

(a) Find the total force acting upon a plane whose area is 0.5 sq. in. and which is perpendicular to the unit vector $\mathbf{u}$, whose components are

$$u_1 = 1/\sqrt{6} \qquad u_2 = 1/\sqrt{6} \qquad u_3 = 2/\sqrt{6}$$

(b) Resolve this force into components in (1) a direction perpendicular to the plane, (2) parallel to the line of intersection of the plane with the x_1x_2 coordinate plane, and (3) a direction perpendicular to the preceding two directions.

3.3 The diffusivity tensor of copper in an anisotropic material has the following components in a principal coordinate system in units of 10^{-12} cm^2/sec:

$$D_{11} = 6 \qquad D_{22} = D_{33} = 3$$

Find the current density of copper at the point (1, 1, 1) if the instantaneous concentration of copper throughout the system is

$$c = (x_1^2 + 3x_2)x_3.$$

3.4 The components of two vectors $\mathbf{A}$ and $\mathbf{B}$ are functionally related by the equations

$$A_1 = 2B_1 + 3B_2 + B_3$$
$$A_2 = -2B_2$$
$$A_3 = B_1 - 3B_2$$

Evaluate the components of the tensors $d\mathbf{A}/d\mathbf{B}$ and $d\mathbf{B}/d\mathbf{A}$.

3.5 A scalar ϕ is functionally related to the components of a vector $\mathbf{A}$. Show that the quantities $\partial\phi/\partial A_i)_{Aj,Ak=\text{cons.}}$ are the components of a vector.

3.6 Let $\mathbf{T}$ be a tensor of rank three, $\mathbf{S}$ a tensor of rank two, and $\mathbf{V}$ a vector, all of which are related by $\mathbf{S} = \mathbf{T}\cdot\mathbf{V}$. If $\mathbf{S}$ is symmetric, how many distinct components does $\mathbf{T}$ have in an arbitrary coordinate system? How many independent components does it have?

3.7 Show that the stress whose components are

$$S_{11} = 3 \qquad S_{33} = 2 \qquad S_{13} = S_{31} = 3$$
$$S_{22} = 2 \qquad S_{12} = S_{21} = 0 \qquad S_{23} = S_{32} = 0$$

(all in pounds per square inch) is equivalent to a hydrostatic pressure of 2 ppsi, an S_{11} tensile stress of 1 ppsi and a shear of 3 ppsi. Note that pressure is really a rank two tensor whose components are invariant under an arbitrary transformation of coordinates.

3.8 Let **T** be a tensor of rank four whose non-zero components are

$$T_{i(ii)} = 5 \qquad i = 1, 2, 3$$
$$T_{i(i)j(j)} = 6 \qquad i, j = 1, 2, 3$$
$$T_{ij(ij)} = 1 \qquad i, j = 1, 2, 3.$$

B and **C** are vectors. The magnitude of **B** is fixed at 10, it lies in the x_1x_2 plane and its direction may vary. The components of **C** are $C_1 = 5$, $C_2 = C_3 = 0$. Evaluate the components of

$$\mathbf{V} = \mathbf{C} \cdot \mathbf{T} : (\mathbf{BB})$$

when **B** forms the angles 0, 30°, 45°, 60°, and 90° with the x_1 axis.

4

STRUCTURE OF SOLIDS

The first three chapters of this text were concerned with the development of the mathematics necessary to provide a phenomenological description of how materials behave. We will also be concerned with a mechanistic explanation of the behavior of solids, and this must be based upon an understanding of the structure of solids.

4.1 Classification of solids

On the basis of structure, materials are divided into two broad categories, *crystalline* and *noncrystalline* or *amorphous*, which are primarily distinguished from one another by the degree of order discernible in the arrangement of the fundamental particles comprising them. In noncrystalline solids the fundamental particles are arranged more or less randomly with respect to one another with only minor elements of order being discernible in the array. In crystalline solids the atoms, ions, or molecules comprising the individual crystal are arranged in a three-dimensional lattice following a precise periodic pattern with relatively minor defects disturbing the general order of the array.

Crystalline solids may consist of one or more single crystals. Until about the late 1940's, polycrystalline materials were of predominant importance in the field of engineering. Since that time a large number of electronic devices based upon nearly perfect single crystals have been developed and now both types of material are of about equal importance. Truly perfect crystals never occur in nature and can be synthesized only with extreme care. However, before we can discuss the imperfections that occur in crystals, it is

necessary to define the structure of ideal crystals; therefore the discussion here is concerned with perfect crystals.

4.2 Crystal structure

The basic cell. As was noted previously, the particles comprising a crystal are arranged in space in a three-dimensional periodic lattice. Since the

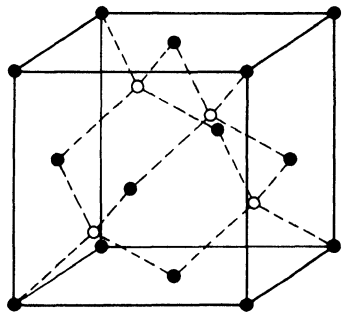


Figure 4.1 ZnS (Zincblende) basic cell.

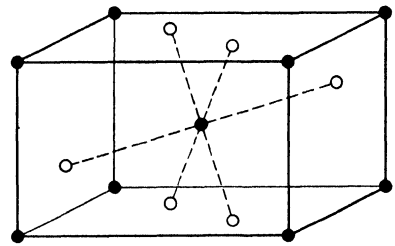


Figure 4.2 The rutile (TiO_2) lattice.

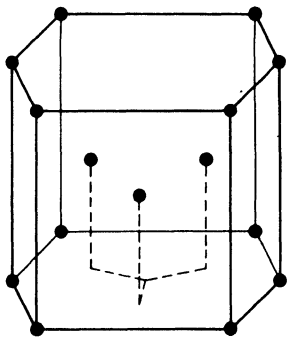


Figure 4.3 The hexagonal close-packed structure.

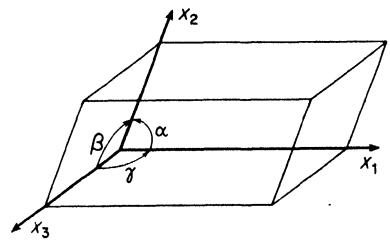


Figure 4.4 The general triclinic prism from which the polyhedra associated with the basic cells are derived.

lattice is periodic in space, it will consist of a set of identical adjacent basic units, each of which by a suitable translation could be brought into coincidence with any other. These fundamental arrays are called *basic cells*. Figures 4.1 through 4.3 show examples of basic cells of three different crystal types.

When drawing these basic cells in perspective, it is quite natural to outline a polyhedron with certain additional guide lines depending upon the complexity of the structure of the basic cell. Although these polyhedra

have been introduced simply as a convenience for showing the location of the atoms in perspective, we shall find that, if they are properly chosen, they provide a useful conceptual device for classifying crystals and for defining certain types of coordinate axes. For many purposes it is easier to describe the geometrical properties of the crystal in terms of the associated polyhedron than in terms of the basic cell itself. The criteria for choosing the associated polyhedron are:

1. that translation parallel to any edge of the polyhedron by a distance equal to the length of that edge will bring the cell into coincidence with an adjacent basic cell, and
2. that such primitive translations be smaller than any other set of translations that would produce coincidence.

In any given lattice there are many different sets of translation distances and directions that could produce a coincidence of basic cells. The second restriction provides the limitation necessary to associate a unique polyhedron with each general type of basic cell.

In reference to the aforementioned figures, we note that an atom located at the corner of the polyhedron is shared among six different basic cells, while atoms on an edge are shared among four and atoms on faces are shared between two basic cells. Not all the atoms of the basic cell are necessarily inside the associated polyhedron. Those located outside are shared by adjacent cells. There is no formal necessity for having an atom located at the corners of the associated polyhedron, although this situation will very often occur.

Crystallographic axes. For many purposes it is convenient to choose a set of coordinate axes defined in terms of the polyhedron associated with the basic cell. The axes of this coordinate system are called the *crystallographic axes.* They are obtained by erecting a set of coordinate axes parallel to the edges of the polyhedron and choosing the distance units along each of the axes to be equal to the corresponding length of the basic cell. In some problems it may be desirable to use a set of Cartesian coordinates with absolute distance units, such as the Ångstrom unit. It is very important to recognize which set of coordinates is being used in any particular discussion. In general, crystallographic axes are convenient for specification of planes and directions in the crystal; but where an absolute computation of length, area, or volume is required, a regular set of Cartesian coordinates is used.

When we have occasion to use a Cartesian coordinate system in conjunction with crystallographic axes, we shall observe a common convention in selecting the orientation of the Cartesian coordinates with respect to the crystallographic axes. It will not be attempted to explain this in detail; instead, in Table 4.1 the conventional orientation of the Cartesian system is

shown superimposed upon the crystallographic axes for each of the six crystal systems.

Directions in the crystal are designated by the components of a vector along the crystallographic axes. Since a vector of arbitrary length will serve to designate a direction, the length is chosen so that its components will be a set of integers having no common factor other than unity. These are distinguished by enclosing the set in brackets ⟨ ⟩.

Miller indices. Once the crystallographic axes have been defined, we may designate planes in terms of their intercepts along these axes. Usually those planes of interest will intersect the crystallographic axes at points occupied by atoms. In any particular crystal there will be many equivalent parallel planes, and the set of these may be designated by a set of numbers called *Miller indices.* These are derived in the following way:

1. Choose one plane of the set of interest and note its intercepts with the crystallographic axes. These must be chosen so that no intercept is at the origin.
2. Form the reciprocals of these intercepts.
3. Convert these to the smallest set of whole numbers that can be obtained by multiplying each of the fractions by the same number.

These sets are enclosed in parentheses to distinguish them. The steps are illustrated in Figs. 4.5 and 4.6, which show a number of different crystallographic planes with their corresponding intercepts and Miller indices for a cubic and an orthorhombic structure. In interpreting a set of Miller indices,

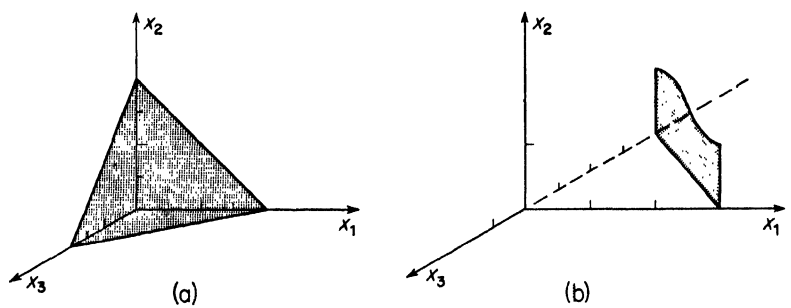


Figure 4.5 (a) One of a set of planes having Miller indices (111)
Intercepts 2,2,2
Reciprocals $\frac{1}{2},\frac{1}{2},\frac{1}{2}$
Reduction to whole No. 1,1,1
(b) One of a set of planes having Miller indices (403)
Intercepts $3, \infty, -4$
Reciprocals $\frac{1}{3}, 0, -\frac{1}{4}$
Reduction to whole no. 4, 0, 3

the first number refers to the intercept on the x_1 axis, the second to the x_2 axis, and the third to the x_3 axis. A line over a number indicates a negative. Since the distance units on the crystallographic axes differ depending upon

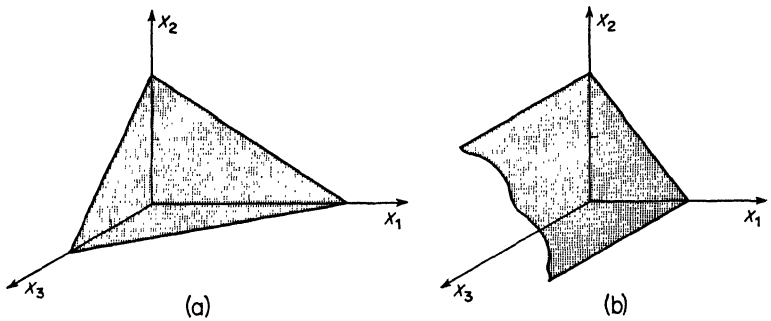
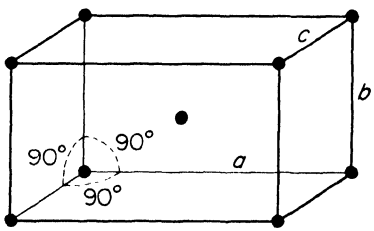


Figure 4.6 Miller indices of certain crystallographic planes in an orthorhombic crystal. (a) (111); (b) (210).

the crystal system under consideration, pairs of planes having the same sets of Miller indices will intersect at different angles in different crystal systems.

Example 4.1. The diagram shows the basic cell of a hypothetical material whose cell dimensions, a, b, c, are in Ångstrom units (1 Å = 10^{-8} cm) and whose atomic weight is M. There are 6.023×10^{23} atoms (or molecules) in one gram molecular weight of the material.



(*a*) Compute the weight density of the substance.

(*b*) Compute the number of atoms per unit area on (110) and (010) planes.

(*c*) Compute the number of atoms per unit length in the $\langle 10\bar{1} \rangle$ and $\langle \bar{1}1\bar{1} \rangle$ directions on the (111) planes.

(*d*) Compute the distances between adjacent (101) planes.

Answers:

(*a*) The atoms at the corners of the basic cell are each shared among eight cells. Therefore, there are only two atoms per cell. Thus the density is

$$\frac{2M}{6.023 \times 10^{23}} \times \frac{1}{abc \times (10^{-8})^3} \quad \text{grams/cc}$$

(*b*) Taking the section of the (110) plane inside the basic cell, its area will be

$$c\sqrt{a^2 + b^2} \quad \text{Å}^2$$

The atoms at the corners are each shared among four other similar plane sections. Therefore, there are 1 (that is in the center) plus $4(\frac{1}{4})$ atoms in this area. The number of atoms/unit area then is

$$\frac{2}{c\sqrt{a^2 \times b^2}} \times 10^{16} \quad \text{atoms/cm}^2$$

In similar manner, the density of atoms on the (010) planes is $10^{16}/ac$ atoms per cm^2.

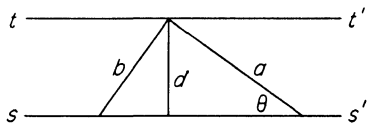
(*c*) The linear density of atoms in the $\langle \bar{1}01 \rangle$ direction on (111) planes is

$$\frac{10^8}{\sqrt{a^2 + c^2}} \quad \text{atoms/cm}$$

and, in the $\langle \bar{1}1\bar{1} \rangle$ directions it is

$$\frac{2 \times 10^8}{\sqrt{a^2 + b^2 + c^2}} \quad \text{atoms/cm}$$

(*d*) The intersection of (101) planes on a (010) plane is shown by the lines ss' and tt'. The distance $d = a \sin \theta$ where $\tan \theta = b/a$.



Crystal systems. The external appearance of a crystal depends primarily upon any constraints that it is subject to during growth, and upon mechanical abuse that it may suffer subsequently. However, the surface of a crystal will always consist of crystallographic planes having low Miller indices, usually ranging from 1 to 4. (See Fig. 4.7.) These planes may only be detectable microscopically, with the macroscopic appearance of the crystal having the form of a continuous contour, but under careful examination the surface will always be found to consist of plane sections. Since these are crystallographic planes, they will always form angles with one another that are characteristic of the way the atoms are arranged in the basic cell. Long before the existence of atoms or molecules had any more experimental justification than a philosopher's hypothesis, this law of rational indices had been discovered and used to classify crystals. Although this classification system was developed around measurements made of the macroscopic features of crystals, it can be applied to the basic cell just as well. It is from this point of view that we shall use it.

In classifying crystals according to their structure, we first define six crystal systems in terms of the shape of the polyhedron associated with the basic cell. Starting with the general polyhedron shown in Fig. 4.4, we may form five of the six polyhedra of Table 4.1 by setting combinations of sides equal to one another and letting two or more of the angles be equal to 90°. A special convention is used in labeling the sides and is incorporated into the

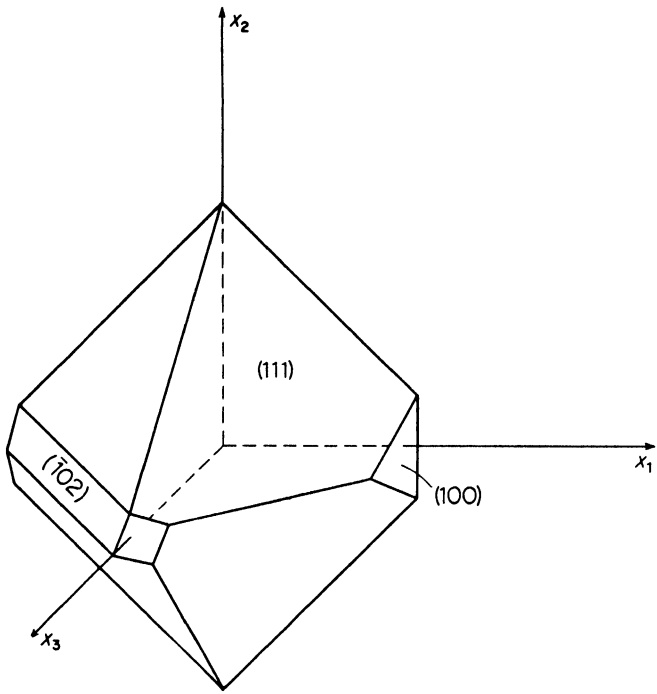


Figure 4.7 Possible external shape of crystal having a cubic structure.

definitions of the systems in Table 4.1. Some authors separate the trigonal-hexagonal system into two separate classes, but these are closely related since a hexagonal or a rhombohedric structure can be synthesized from a grouping of trigonal prisms in which the interior angles of the bases are 60° and 120°.

Within each of these systems there are a variety of ways that atoms may be arranged and still be represented by the same polyhedron. From the very simplest structures in which an atom is located at the corners of the polyhedron, we may increase the degree of complexity slightly by forming face- or body-centered modifications. (See Fig. 4.8.) The face-centered and body-centered cubic structures as well as the hexagonal close-packed structure are of special interest since over 70 per cent of the known metals and alloys crystallize in one of these forms.

Table 4.1 *Crystal systems*

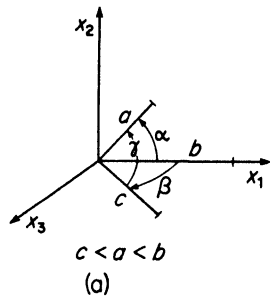
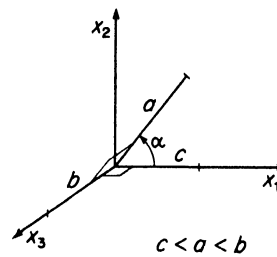
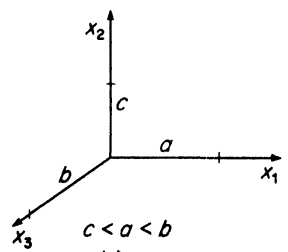
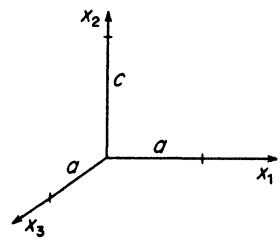
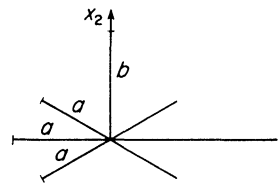
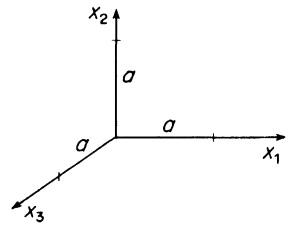
System	Definition	Crystallographic axes
I. Triclinic	Three sides at oblique angles to one another	 (a)
II. Monoclinic	Three unequal sides, one side perpendicular to the plane of the other two	 (b)
III. Orthorhombic	Three unequal mutually perpendicular sides	 (c)
IV. Tetragonal	Three mutually perpendicular sides, two of equal length	 (d)

Table 4.1 *Crystal systems (continued)*

System	Definition	Crystallographic axes
V. Trigonal-hexagonal	Three axes of equal length in one plane, perpendicular to a fourth axis	 All angles = 60° (e)
VI. Cubic	Three equal mutually perpendicular sides	 (f)

Coordination numbers. Frequently, the number of nearest neighbors each atom has in a particular structure provides an indication of the density of the structure and also a clue to the type of bonding between atoms. This number is called the *coordination number*. If all of the atoms are of the same size and are spherical, geometrical considerations dictate a maximum coordination number of 12. The face-centered cubic and hexagonal close-packed structures are the only ones possible having a coordination number this high.

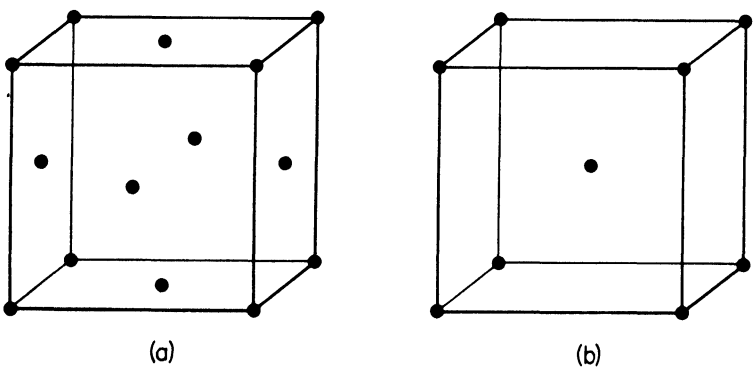


Figure 4.8 (a) Face centered cubic lattice. (b) Body centered cubic lattice.

The next highest coordination number 8 occurs in the bcc lattice. Approximately 70% of all elementary metals and alloys have either the fcc, hcp, or bcc structure. This is one reason for the high density of metals in comparison with nonmetallic solids.

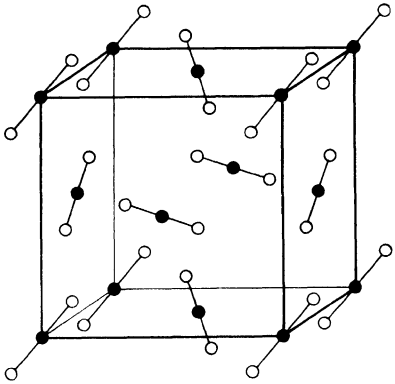


Figure 4.9 The pyrites (FeS_2) structure.

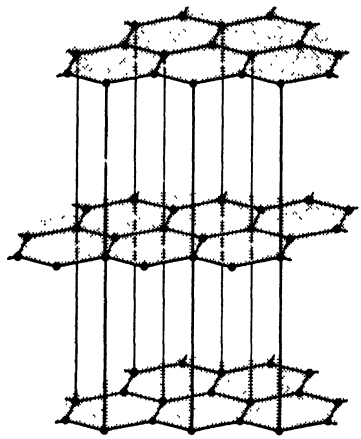


Figure 4.10 The graphite lattice.

Figures 4.9 and 4.10 show some structures of somewhat greater complexity than any considered previously. These are included primarily as an illustration of some of the different ways in which atoms may be arranged in conjunction with a simple polyhedron. In general, the more complex the structure of the basic cell, the lower its symmetry.

4.3 Crystal defects

It was noted earlier that the discussion was strictly applicable only to ideal crystals which were free of defects. These defects, when present in normal concentrations, may have little or no effect upon some crystal properties. However, they do have a significant effect upon optical and electrical properties, and in the case of the mechanical properties of crystals the defects play a dominant role. Except insofar as it may be helpful in making the discussion clearer, we shall not attempt in this section to show the relations between defects and their effects upon properties, but will direct our major efforts toward describing and characterizing them. In later chapters, particularly those concerned with mechanical and electrical properties of solids, we shall show the mechanisms by which defects affect properties.

Dislocations. Of all the types of crystal defects known to exist, dislocations are perhaps the most versatile in explaining what would otherwise be anoma-

lous behavior in the mechanical properties of materials. Dislocations also enter into the mechanics of crystal growth in a crucial way. As a consequence nearly all crystals, except those grown under very special conditions, will have a rather high concentration of dislocations. We shall concern ourselves here with three types of dislocation: the edge dislocation, the screw or spiral dislocation, and the semidislocation. The lattice defect arising from a missing atom in the lattice will be considered to be a special case of the semidislocation.

The *edge dislocation* is the easiest of all to visualize. Figure 4.11 shows the lattice points of a portion of a (111) plane of atoms in a perfect face-centered cubic crystal. Now imagine that those atoms on the sites marked with an x

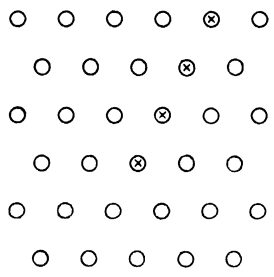


Figure 4.11 Lattice points in a portion of the (111) plane of a face-centered cubic crystal.

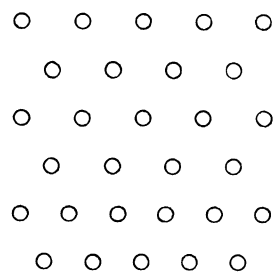


Figure 4.12 Same lattice as in Fig. 4.11 after the lattice points marked by an x were removed.

in Fig. 4.11 are missing as well as all the rest of the atoms on the same line from the upper righthand x on out to the surface of the crystal. Then the atoms on each side of the line of vacancies and the crystal lattice would be deformed in the manner shown in Fig. 4.12. Note that the strongest deformation of the lattice is in the neighborhood of the point in Fig. 4.12 corresponding to that occupied by the lowest x in Fig. 4.11. In moving away from this point in any direction the lattice strain decreases uniformly, until after five to ten interatomic distances the lattice assumes its normal form. This is a two-dimensional schematic representation of an edge dislocation. It is important to realize that in two dimensions the dislocation is not the line of missing atoms, but rather the point at which the line of missing atoms terminates. In a three-dimensional model of a crystal an edge dislocation would be formed by the removal of the atoms residing upon a portion of a plane extending from the crystal surfaces down into the crystal. Then the dislocation would consist of a line or curve along the edge of the plane of missing atoms.

Figure 4.13 is a schematic drawing of a crystal with the shaded semiplane showing the locus of the lattice vacancies. The dislocation is a *line defect* located along the line aa' in the figure. In a purely geometric sense, that

portion of the plane terminating inside the crystal could be a line forming any angle with the other edges of the plane or could even be a curve. These general ideas could be specialized to include dislocations formed by lattice vacancies on a portion of a plane entirely inside the crystal as shown in Fig. 4.14, in which case the defect would be a closed curve. This is called a

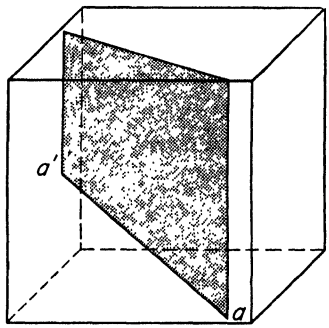


Figure 4.13 Shaded plate represents the portion of the crystallographic plane from which the atoms are missing.

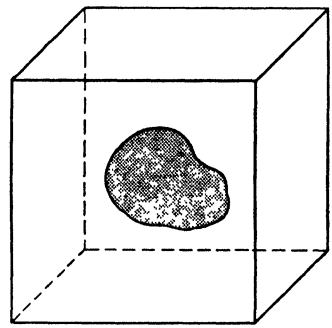


Figure 4.14 The semi-dislocation is a closed curve defect around a portion of a plane entirely inside a crystal from which all of the atoms are missing.

semidislocation. The plane could also include just one missing atom, and the dislocation would be a *point defect*.

An entirely different type of dislocation called a *spiral dislocation* is often formed in the process of crystal growth. The structure of such a dislocation can best be described by imagining that a cut is made partway through a crystal along a crystallographic plane, and that on one side of the cut the crystallographic plane is shifted upward or downward by one or more unit distances as shown in Fig. 4.15. The dislocation itself will consist of a strain field localized along the line marking the terminus of the cut. The name "spiral" is given to this type of dislocation because the originally parallel crystallographic planes have been partially sectioned and displaced to form a single continuous helicoid. This may be easier to visualize if we represent the crystallographic planes by several sheets of paper stacked on top of one another. If we cut the stack partway through and joined the part of sheet 1 on the left of the cut to the part of sheet 2 on the right of the cut, and proceeded through the stack in this manner, we would end up with one single sheet forming a continuous helicoid.

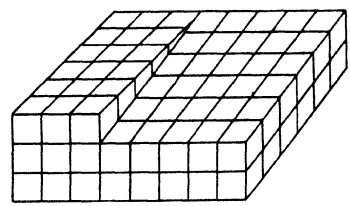


Figure 4.15 A spiral dislocation.

The resolution of mechanical anomalies by dislocation theory depends

upon the fact that dislocations are mobile and will migrate when a strain field is induced by an external stress. The mechanism by which an edge dislocation moves in a strain field is illustrated in Figs. 4.16a to 4.16c. The z axis of the coordinate system in the figures is directed outward from the plane of the paper. If a stress is applied with an xy component, the atoms above the

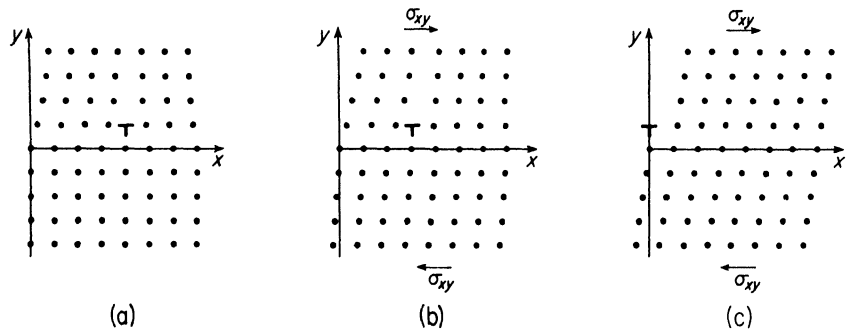


Figure 4.16 The motion of an edge dislocation in a strain field. Progressively greater strain will drive the dislocation to the edge of the crystal plane.

x axis will be displaced to the right and those below the x axis will be displaced to the left. A strain consisting of a shift of one plane of atoms with respect to another by a small fraction of an interatomic distance will cause the dislocation to migrate through one or more interatomic distances. Figures 4.16a through 4.16c also show progressive stages of the motion of an edge dislocation due to an external stress. Certain of the atoms have been numbered to show the motion of the dislocation relative to the lattice. A reversal of the sign of the strain component would have caused a reversal in the direction of the motion. Consequently, if a dislocation were moving through a crystal under the influence of a strain from an external source, and should move into a region of inhomogeneous strain having a component opposite to that from the external stress, its motion would be inhibited.

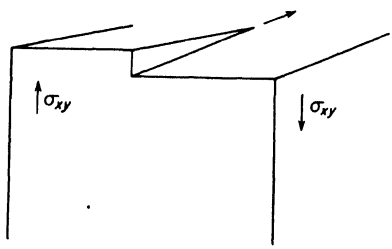


Figure 4.17 Motion of spiral dislocation inducing a shearing stress.

The mechanism by which a spiral dislocation can move under the influence of a shear stress is somewhat easier to understand than that of the edge dislocation. Under the influence of a shear stress the spiral dislocation of Fig. 4.17 would move in the direction shown, just as a tear would propagate through a piece of paper under a shear stress.

Just as the motion of dislocations plays an important role in the explanation of certain aspects of the mechanical behavior of solids, those factors which inhibit the motion of the dislocations must play an equally important role. Anything that is capable of inducing a local inhomogeneous strain in the lattice is capable of inhibiting the motion of dislocations. The list of these factors would include other dislocations, inclusions of impurities, twin planes, grain boundaries, and misplaced atoms. Two dislocations that are properly oriented with respect to one another may also partially or completely annihilate each other.

When a dislocation moves through a lattice and encounters an inhibiting strain field, the strain field will not normally extend all along the dislocation line. Thus, the dislocation would not become immobilized along its entire length, but more likely would be caught at one point. Under these circumstances the dislocation line would bend and under sufficient external stress would eventually break loose from the inhibiting strain field, and in so doing it would generate more dislocations in its wake. The net effect of an external stress is to cause some dislocations to migrate to the surface and out of the crystal as well as to generate new dislocations by the mechanism mentioned above. They also may be generated at points of high local stress.

Almost all of the characteristics that we have ascribed to dislocations in the previous paragraphs may be illustrated by using a bubble raft as a macroscopic analogue of the atoms on a plane of a crystal. This includes the motion of dislocations due to strains caused by external stress, the immobilizations by odd-sized bubbles (representing impurity atoms), the generation of more dislocations by forcing dislocations past inhibiting strain fields, and the generation of dislocations at points of high local stress. To one who has seen a motion picture of such a bubble raft, words become relatively ineffective devices for describing the structure and properties of dislocations. Figure 4.18 shows a picture of a bubble raft with a dislocation in it. The location of the dislocation may be best found by looking at the illustration from an acute angle.

The role played by dislocations in crystal growth can best be understood by reference to Fig. 4.17. The spiral dislocation provides a partial step at the surface. When an atom strikes the surface of a growing crystal, the probability of its staying on the surface is increased if it strikes at a point where it can form bonds on two or more sides rather than one. This has a striking effect upon the crystal growth rate and is the only mechanism so far proposed which is capable of explaining the observed rate of crystal growth. With the aid of the electron microscope, direct evidence has been obtained for this type of growth mechanism in the form of growth spirals which may be observed on crystal surfaces. (See Figure 4.19.)

The special type of dislocation which consists of a single lattice vacancy produces a fairly symmetric strain field, and is particularly important in the

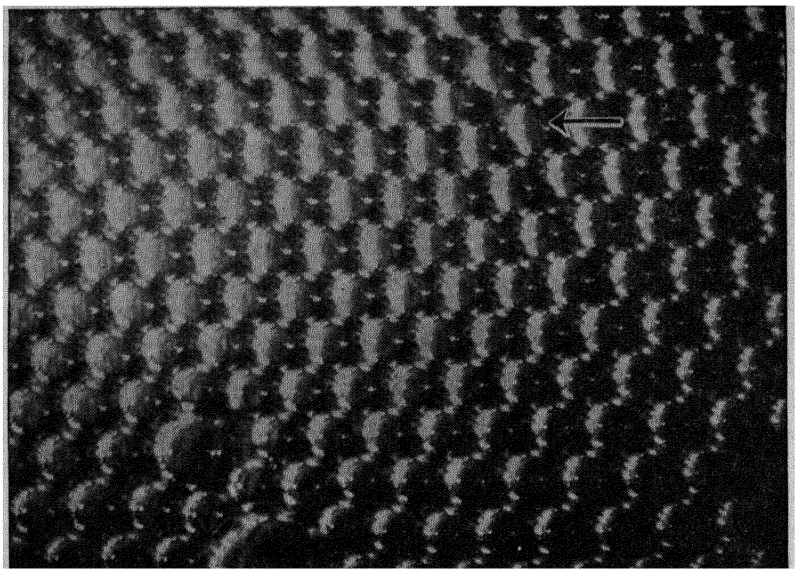


Figure 4.18 Bubble raft model of an edge dislocation.

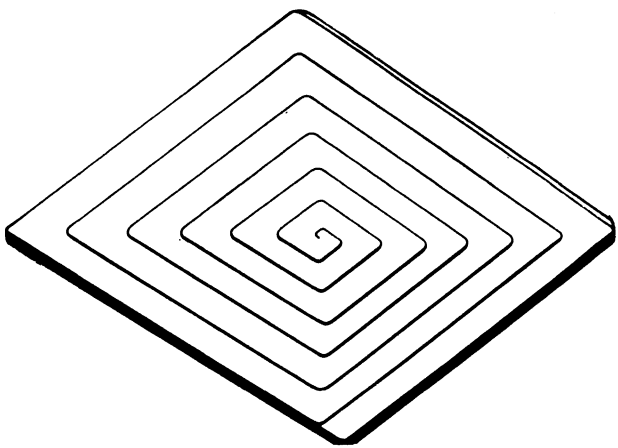


Figure 4.19 Tracing of a photo-micrograph of a spiral growth ring on a crystal of paraffin.

transport of impurities through solids. Often such a lattice vacancy is found in conjunction with a misplaced atom which is located at a point in the crystal where there is no regular lattice point available for it to occupy and which is accommodated only by a severe distortion of the surrounding lattice. Such pairs of defects may be formed by radiation sufficiently energetic to cause an atom to be ejected from its normal site. This is one of the forms of radiation

damage which occurs in solids subjected to a high neutron flux in reactors. Often it may be manifested by the discoloration of ionic crystals which have been irradiated with X-rays. The discoloration is due to color centers which are produced by the entrapment of electrons in lattice vacancies previously occupied by a negative ion. Usually the discoloration produced in this manner can be removed by heating the crystal. This increases the mobility of misplaced ions and they may more readily diffuse to a normal lattice site.

Impurities and alloys. Although it may be stretching our semantics a bit to call the inclusion of an impurity atom in the lattice a crystal defect, the normal consequence of such an inclusion will be a strain field in the lattice around the impurity site. This is explained by the fact that even if a foreign atom should be of such a physical size that it would not perturb the lattice, the bonds which it would form with its neighbors would differ to varying degrees from those which were normally formed between atoms of the same kind in the lattice. This would tend to distort the electron clouds of the neighboring atoms and would set up local strain fields.

There are actually two ways in which an impurity atom may enter into the lattice of a crystal. A substitutional impurity simply occupies a lattice site which would normally have been occupied by an atom of the kind of which the crystal is mainly comprised. However, if the inpurity is sufficiently small with respect to the atoms that are the major constituents of the crystal, the impurity may reside at the interstices between the other atoms in the lattice. Even when the impurity is small enough to be accommodated at such a point geometrically, it will still perturb the electronic states of its neighbors thus tending to change their size and setting up strain fields in the lattice.

4.4 Polycrystalline materials

Grain boundaries. Most of the common materials of engineering interest are not single crystals, but consist of a large number of crystals bonded to one another along interfaces called *grain boundaries.* The structure of a typical polycrystalline metal is illustrated by the photomicrograph of Fig. 4.20. The average crystal size has an important effect upon mechanical properties and, to a lesser degree, crystal size affects properties such as diffusivity, electrical and thermal conductivity, and yield point. These effects are largely attributable to the action of the grain boundaries. In order to explain later their effect upon specific properties, it is desirable to consider their structure in some detail here.

The presence of impurities will usually make the structure of the grain boundary somewhat more complicated, so let us first consider the structure of grain boundaries in pure materials. A cursory glance at Fig. 4.20 would seem to make the definition of a grain boundary rather self-evident; it is an

interface between two adjacent crystals. However, if two crystals were placed in contact with one another with a perfect alignment of corresponding crystallographic axes, the surface atoms would bond together and the two crystals would fuse into a single crystal. Consequently, it is necessary that there be a change in the direction of the crystallographic axes in crossing a grain boundary. Immediately this puts us into some semantic difficulties,

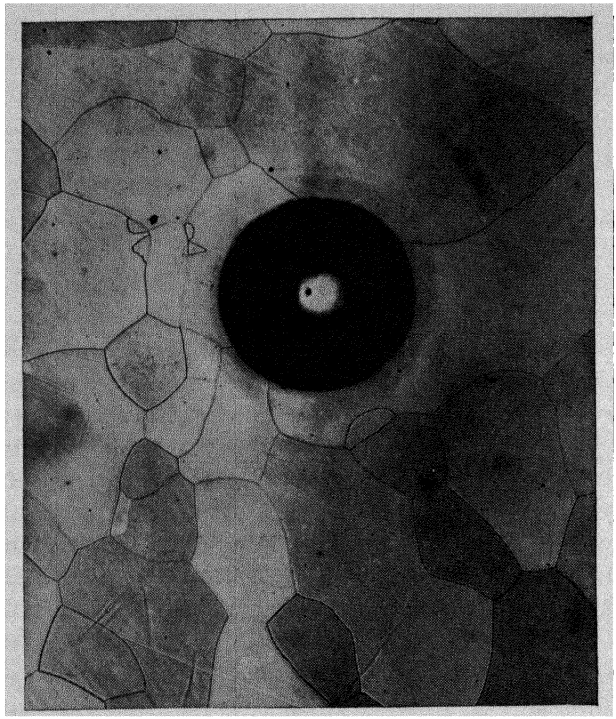


Figure 4.20 Photomicrograph of the grain boundaries in an Fe-Al alloy. (Photograph courtesy of R. W. Cahn, Department of Physical Metallurgy, University of Birmingham.)

because by this definition any very small nonzero change in the direction of the axes must be considered a grain boundary. Actually very small changes in the directions of the axes may be present without appearing on micrographic examination and without having an appreciable effect upon the properties. When such grain boundaries occur, we do not usually consider them as boundaries between different crystals, but consider the region containing them as one crystal which is polygonized or divided into subgrains. The formation of subgrains in a single crystal is usually the result of mechanical stress and is detectable by the diffuse background it gives to the X-ray diffraction pattern of the crystal. A highly polygonized single crystal may

have such a diffuse X-ray diffraction pattern that it will resemble that of an amorphous material rather than a single crystal.

When the angle of misfit between the axes of adjacent crystals approaches one or two degrees, the grain boundary begins to assume the character of a true interface between adjacent crystals. As viewed from inside the crystal, the atoms on the interface are not bonded as symmetrically as those inside the crystal. This produces strain fields near the interfaces which decrease in intensity as one moves away from the interface. These strain fields represent stored energy which is identical with the surface energy due to surface tension at any interface. This strain energy is not only a function of the angle between the axes of adjacent crystals, but also depends upon the plane of the angle. The effect of the grain boundary upon material properties is largely determined by the magnitude and extent of the strain field associated with the boundary. The strain field is greatest when the axes on each side of the boundary form high angles and have a skew orientation with respect to one another. The effect upon the mechanical properties of a pure material is largely due to the inhibiting action of the strain fields upon the motion of dislocations. The electrical properties are affected by the scattering of charge carriers by the strain fields and a warping of the energy states in which the charge carriers must reside. The high degree of disorder along a grain boundary naturally is more permeable to diffusing material, and the extra strain energy in the immediate neighborhood of a boundary increases the chemical activity locally. Thus, corrosion proceeds more rapidly at grain boundaries, and when the material is undergoing a phase change, the boundaries provide the energy for the nucleation of crystals having the structure of the phase into which the material is changing.

Despite the extensive disruption of the periodicity of the structure at a grain boundary, the atoms will bond across grain boundaries. Since these bonds are distributed in space in an abnormal manner, and are not of the usual bond length, they will not be as strong as those inside the crystal. For this reason the ultimate tensile strength of a polycrystalline material may be lower than that of a single crystal. However, the yield point of a polycrystalline material will ordinarily be 1000 per cent higher than that of a single crystal of the same material because of the lower dislocation mobility. Also the network of internal stresses will increase the hardness of a polycrystalline material.

Twinning. Figure 4.21 shows a two-dimensional sketch of the location of the atoms along the boundary between the two parts of a crystal twin. It is worth noting that the two parts of the crystal on each side of the twinning plane are mirror images, that there is an abrupt change in the orientation of the crystallographic axes across the twinning plane, and that every atom is in a proper lattice site. The lattice points on the twinning plane are shared by

the two parts of the crystal. The angle of misfit along a twin boundary will always correspond to the angle between planes in which the Miller indices consist of two like indices with the third differing by no more than one or two. Despite the high angle of misfit, the accompanying strain energy will be relatively low. The reason is that one crystallographic axis on one side of the twin will be parallel to the corresponding axis on the other side of the plane. To state the situation more explicitly, we might label the crystallographic axes to the left of the twin boundary x_i and to the right y_i. Then in order to have a twin boundary, two pairs of axes, let us say x_1 and y_1, x_2 and y_2, must be oriented at specific angles with respect to one other, while x_3 and y_3 remain parallel. Under these circumstances, the (010) planes would meet along a line. If the specific angles required for twinning are not present, but the x_3 and y_3 axes are parallel, the (010) planes would still meet along a line, and the boundary would consist of a series of edge dislocations whose density would depend upon the angle between the other axes. However, in either of these cases, there would be considerably less disorder and, consequently, less strain energy involved in the boundary than there is when none of the axes on either side of the boundary is parallel to an axis on the other side.

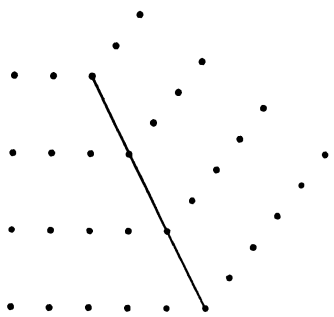


Figure 4.21 Twin boundary.

Twins may be formed by plastic deformation of the crystal. If conditions subsequent to the original deformation are such that they promote crystal growth, the twinned crystals may grow to be fairly large. Metals that have been rolled and subsequently annealed may show rather large twinned crystals which are easily visible. Frequently the galvanizing coating on steel will also show large numbers of twins. Quartz crystals that are to be used as piezoelectric devices must be selected so that there are no twin boundaries, since the piezoelectric effects on each side of the twin boundary would cancel one another.

Segregation. When an impurity is present or when a material is deliberately composed of two or more constituents, the structure and properties of the grain boundaries may become somewhat more complex. Figure 4.22 shows a phase equilibrium diagram for two hypothetical elements, *A* and *B*. If we begin with *A* as the major constituent and it is contaminated with 1% *B*, then heating the mixture to a temperature corresponding to point *a* on the phase diagram will bring the entire mixture into a liquid state. If the mixture is cooled to point *b*, then solid crystals of composition *c* will separate out. If the mixture were then cooled infinitely slowly, the composition of the solid

phase would follow along the curve *fg* and the composition of the liquid would follow *fh*. The composition of the solid phase cannot change very rapidly and if the mixture were cooled at a reasonable rate, crystals of composition *c* would first separate out and commence to grow, with the liquid becoming progressively richer in constituent *B*. At the temperature *g*, the mixture would consist of crystals having a composition of about that corresponding to point *c* surrounded by a film of liquid whose composition would be about that designated by point *h*. Further cooling would cause all of the

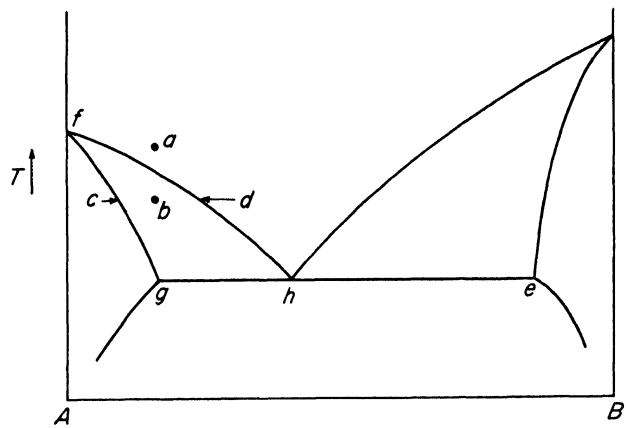


Figure 4.22 Phase equilibrium diagram of two hypothetical materials.

liquid to disappear, leaving crystals of *c* with material of approximately the composition *e* embedded along the grain boundaries.

This is called *segregation*, and just as it seems undesirable in human society, it may also be disadvantageous from a metallurgical point of view. However, in this context, segregation is not always bad. Hypoeutectoid steels, for example, consist of crystals of almost pure iron embedded in a matrix of segregated pearlite. The latter provides the structure with rigidity, while the softer iron core provides resilience. More frequently, segregation along grain boundaries may give rise to a host of undesirable characteristics. The most obvious of these is that the segregated materials have a lower melting point than the major constituent. If there is an extreme difference in melting point, the crystals will simply fall apart under very light loads at whatever temperature the segregated phase becomes liquid. Aside from this factor, the mechanical properties of the segregated phase may be rather poor and this would be reflected in the cohesiveness of the entire matrix of crystals. The segregated material is normally crystalline itself, but when it is constrained to solidify in the crowded region between the growing crystals of the major constituent, it is likely to have such a highly

disturbed structure that its mechanical properties will be poor—even if they are fairly good when it crystallizes normally.

Orientation. Aside from the influence of the presence of grain boundaries in a polycrystalline material, the relative orientations of the crystallographic axes of the individual crystals comprising the material also have an effect upon the properties of the sample, but in an essentially different way. The individual crystals will be anisotropic in at least some of their properties, and unless this anisotropy is canceled by a random distribution in the orientation of the crystallographic axes it will carry through to the properties of the polycrystalline sample. The exact relation between structural symmetry and anisotropy will be explored in detail in Chap. 7, but in general, the more symmetric the structure, the fewer of its properties will exhibit anisotropy. The structure of a polycrystalline material will always have a symmetry as high as or higher than the symmetry of its constituent crystals. Consequently, a bias in the orientation of the crystallographic axes of a polycrystalline material whose crystals belong to the cubic system is less likely to lead to anisotropy than would be the case if the individual crystals were members of the hexagonal system or some system of still lower symmetry.

The presence of a preferred orientation in the crystallographic axes arises from the very first step in the fabrication of a polycrystalline material. When a crystalline material is cooled from the melt, there is a strong tendency for the crystals to nucleate and grow with one of the crystallographic axes aligned with the temperature gradient at the time of solidification. Thus a cast rod will almost always exhibit a cylindrical symmetry in its grain structure. Subsequent mechanical working such as rolling, forging, or drawing will tend to align the crystals with the direction of the strains induced by the working process. Consequently, a preferred orientation is not an exceptional occurrence, but rather is to be expected and circumvented when necessary. Often an appropriate type of heat treatment will serve to render a material isotropic, but an improperly chosen heat cycle may make it more anisotropic than it was originally.

There are special cases in which anisotropy is desirable. In the case of the laminae used in transformers, fabrication methods are deliberately chosen so that the crystals are so oriented that the directions of highest permeability in the crystal will be nearly parallel to the edges of the laminae.

4.5 The structure of amorphous solids

Although amorphous solids are lacking in the high degree of structural order characteristic of crystals, they may have elements of structural symmetry which will make them anisotropic in their properties. Aside from the anisotropy of their properties, we may later seek explanation of their observed

mechanical behavior, and an understanding of their structure will be helpful in the formulation of such explanations.

Those materials which have the highest structural symmetry are those which have the lowest structural order. When the distances and directions from one atom to another are randomly distributed on a microscopic scale, then on a macroscopic scale the material will seem to possess spherical symmetry and its properties will not vary with direction. The amorphous materials that approach this condition most nearly are those which we call *glasses.*

Glasses. The term "glass" as used here will refer to any supercooled liquid, and although it includes those materials which are normally recognized as glasses, it could just as well include a supercooled solution of glycerine in water. Many organic materials that are highly soluble in water will form glasses when cooled. This is also characteristic of mixtures of many different inorganic oxides. Silica, although it is a normal constituent of many glasses of technological importance, is not a necessary constituent of all glasses.

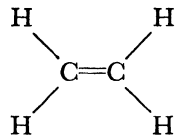
The mechanics of glass formation are as follows. If a molten material is cooled below its normal melting point, it becomes thermodynamically unstable with respect to a transformation into the solid state. In order to undergo such a transformation, crystals must first be nucleated and then must grow by a process that involves the diffusion of the atoms from the liquid phase onto the face of the growing nucleus. The rate of diffusion is limited by the viscosity of the liquid, which in turn usually increases exponentially with the reciprocal of the absolute temperature. Thus if a material is cooled rapidly from the melt, there is always a possibility that the viscosity of the melt will increase so rapidly with decreasing temperature that crystal growth is inhibited. If the inhibition of crystal growth is strong enough, then the material will not crystallize at all, but will remain in a liquid phase. When the temperature has dropped to a value where the viscosity has become so high that the material is essentially a solid, it is called a *glass.* Any glass may be caused to revert to the crystalline state by holding it at a temperature just below the melting point of its constituents for a long time.

Polymers. Amorphous materials such as rubber, wood, polyethylene, and other common plastics have varying degrees of structural order, but the degree of disorder present arises from causes fundamentally different from those which lead to a glasslike structure. In ordinary crystalline materials the atoms or molecules either are spherical or have some other definite, rather compact shape. The molecules present in polymeric systems have a very high molecular weight and are composed of basic units which are repeated at more or less regular intervals. These basic units may rotate about bonds joining adjacent units, and any particular molecule may assume a wide

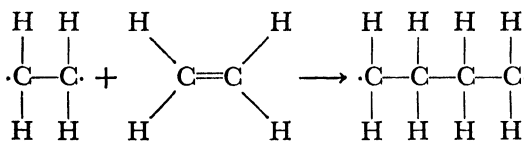
variety of different shapes. Because the molecules of a polymer lack any unique shape and because different molecules may become entangled with one another, they will not be able to crystallize from a melt or from a solution. These factors may be better understood after a discussion of the structure of polymeric molecules, which in turn is best understood in terms of the way polymers are synthesized.

The great majority of polymers either encountered in nature or synthesized are built of chains of carbon atoms joined to one another, with a wide variety of different functional groups attached to those bonds which are not involved in a carbon-carbon linkage along the primary chain. The polymerization of the ethylene molecule will serve as a simple illustration of the polymerization process, and the structure of the resulting molecule will have many of the features common to all polymer molecules.

The ethylene molecule consists of two carbon atoms joined by a double bond, the remaining bonds of each carbon atom being attached to hydrogen molecules. This structure may be represented schematically in the following way:



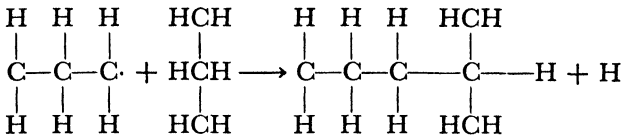
Either thermal excitation or a quantum of energy from some other source may cause an ethylene molecule to assume a temporary structure in which one of the two bonds joining the carbon atoms is broken. While the molecule is in this state, it will be highly reactive, and upon collision with another ethylene molecule a reaction of the following type will occur:



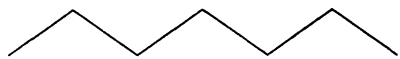
The product of this reaction is also highly reactive because of the unsatisfied valence bonds on the terminal carbon atoms, and it in turn will react with a third ethylene molecule. This process will continue until a long molecular chain has been built consisting of tens, hundreds, or thousands of carbon atoms linked together.

In addition to the primary polymerization process described above, a number of side reactions may also occur simultaneously. If the reactive end of the growing molecule should collide with an impurity such as water or oxygen, the reaction chain would terminate. The reactive terminus of a molecular chain may also react with another polyethylene molecule to form

a branched chain:



The covalent bonds of a carbon atom are directed in space at tetrahedral angles. Taking the simple case of an unbranched molecule in which only the carbon atoms are shown, the molecule would have the following configuration if the chain were lying in a plane:



However, the carbon-carbon bonds offer little resistance to rotation, and there is no a priori reason why the chain should lie in a plane. Instead, each carbon atom might rotate through an arbitrary angle about the bond joining it to the preceding atom, in which case the polymer molecule could assume an infinite number of shapes in three dimensions. Furthermore, the shape would be changing continually under thermal agitation.

The pattern that is beginning to emerge from this discussion has the following salient points. When polymerization is initiated, a large number of different molecular chains will begin to form. These will terminate randomly at different stages of their growth, giving the polymer a distribution of molecular weights. Branched chains may form, and without straining any bonds, the long chainlike molecules may become entangled with themselves and with each other. Superimposed upon the general chaos will be a variety of different modes of molecular motion due to thermal energy.

Polyethylene may be melted without decomposition and may also be dissolved in certain solvents. However, when it does undergo a transition from a liquid to the solid state, the disorder of the liquid state will persist to a large degree—for the simple reason that the decrease in energy of the system that would ordinarily accompany crystallization is not sufficient to disentangle the molecules and allow them to rearrange themselves into an orderly array. Nevertheless, there will be restricted local regions in which a degree of order approaching that of a crystal will exist. Thus polymers will have a certain degree of crystallinity. Although thermal energy is itself a factor in producing disorder, it will also make the molecules more mobile and will help in the rearrangement leading to crystallinity. There will be some temperature at which these opposing factors balance one another; annealing the material at this temperature will maximize the crystallinity of the polymer. Since the regions of crystallinity are usually randomly oriented, they will not ordinarily reduce the structural symmetry of the polymer.

The mechanical properties of plastic films are improved by giving them a two-way stretch after they are extruded. This tends to align the molecules with the direction of stretch, which increases the order and decreases the symmetry of the structure. Materials that have been oriented in this manner will be anistropic. Further modification of the structure may be accomplished during the polymerization by forming bonds between adjacent molecules through the action of heat and an appropriate agent such as sulfur in the vulcanization of rubber. If carried beyond a certain point, cross-linking will produce a hard, infusible mass. Cross-linking may also be accomplished by subjecting the plastic to a flux of high-energy radiation. The usual effect of a radiation flux will be to cause an initial improvement in the mechanical properties of a plastic, but if it is carried beyond a certain point they will eventually begin to degrade.

PROBLEMS

4.1 (a) Calculate the distance between (100) planes and (110) planes in an orthorhombic structure of $c = 3$Å, $a = 2$Å, and $b = 1$Å.
(b) Show that a vector in the $\langle 111 \rangle$ direction is perpendicular to the (111) plane in the cubic system.
(c) For a simple body-centered tetragonal structure, calculate the numbers of atoms per unit area on the (110) planes if $a = 1$Å, $c = 2$Å.
(d) Calculate the linear density of atoms in the $\langle 111 \rangle$ direction in a bcc lattice whose basic cell has a length a on an edge.

4.2 If a bcc structure is made up of rigid spheres, calculate the diameter of the largest sphere which would fit into an interstitial site.

4.3 Show that the diamond lattice consists of two fcc lattices displaced relative to one another by one quarter of the distance along a body diagonal.

4.4 The basic cell of the diamond lattice is 3.56Å on an edge.
(a) Compute the atomic radii of the carbon atoms assuming them to consist of hard spheres that are tangent along tetrahedral lines between each atom and its four nearest neighbors.
(b) If there were a close-packed modification of diamond in which the atomic radii were unchanged, what would be the ratio of the densities of the two structures?

4.5 Compare the structure of an edge dislocation with that of a spiral dislocation.

4.6 Discuss the relationship between the orientation of an edge dislocation and the orientation of the shear strain component which would move the dislocation most readily. Compare this with a spiral dislocation.

4.7 Sketch the arrangement of the atoms on each side of a twin boundary in an fcc lattice where two of the corresponding crystallographic axes on each side of the

boundary are displaced in relation to one another by 45°, whereas the third pair of corresponding axes remain parallel.

4.8 Show that two other types of twin can be formed in the fcc structure in the following way: the crystallographic axes on each side of the boundary share the same ⟨111⟩ direction, but are rotated relative to one another about an axis in the ⟨111⟩ direction through angles of 60° and 180°.

5

BONDING IN SOLIDS

It is not possible to discuss bonding intelligently without using quantum mechanical concepts. To those whose educational backgrounds have not included some contact with this field, quantum mechanics will often seem more like metaphysics than physics. Actually nothing could be further from the truth, but the illusion is more likely to be heightened than dispelled by textbooks such as this in which it is necessary to use many concepts taken from quantum mechanics without providing a proper and logical foundation.

5.1 Introduction

A greater appreciation of the need for quantum mechanics may be gained by considering the nature and significance of the revolution that seized the field of physics about the turn of the century. The classical physics of that time first began to encounter difficulty in problems involving light. Actually a controversy over the nature of light had been raging from the time of Newton and Huygens. This dispute died out in the latter half of the nineteenth century, however, when the wave theory gained ascendancy over the corpuscular theory. Even so, the philosophical difficulty of finding a medium for the propagation of light waves had always been troublesome. Finally the famous Michelson-Morley experiments proved conclusively that no such medium exists. However, a more striking difficulty arose out of Einstein's work with the photoelectric effect in 1905. The result of these experiments was that the energy of the electrons ejected from a metal surface by an impinging light ray was determined by the frequency of the light rather than by its intensity, as was predicted from classical physics. Meanwhile classical

physics was also unable to provide a single unified theory to account for the observed spectral distribution of the radiation from a black-body radiator. Instead there were two theories, one of which worked fairly well in describing the long-wave end of the spectrum, whereas the other worked fairly well on the short-wave end.

Not all of the difficulties were in the area of electromagnetic radiation. Rutherford's scattering experiments in 1916 showed conclusively that the negative components of atoms were distributed outside a positively charged nucleus. According to classical physics, such a system would be unstable and capable of only a transient existence. Finally, and perhaps most significantly, classical physics was completely unable to account for the existence of atomic spectra.

The original introduction of quantum concepts was forced and artificial, and only was justified by the shotgun marriage it effected between theory and experiment. Max Planck developed a theory of black-body radiation which accounted exactly for the observed spectral distribution of a black-body radiator. In order to do so it was necessary to introduce certain postulates which at that time could only be considered as highly artificial. One of these was that the oscillators in a radiating solid could not oscillate over a continuous range of frequencies but only at specific frequencies which were integral multiples of a very small base frequency. In other words, the frequencies of oscillation were quantized.

The major introduction of quantum concepts came about through Bohr's model of the atom. For better or worse, this is usually given full play in even elementary texts, and probably needs little explanation here. It is a highly mechanical model in which electrons are visualized as moving about the atomic nucleus in precise orbits much like a miniature solar system. It provided a good account of the gross features of the spectra of hydrogenlike atoms, and subsequent refinements by Sommerfelt were able to account for the fine structure of atomic spectra. However, in constructing this theory, it was necessary to introduce one postulate that had no counterpart in classical mechanics at all, and another postulate that was at complete variance with a well-established point of classical electrodynamics. The first of these was that the electron orbits could not have arbitrary momenta but had only those orbits in which the angular momentum was an integral multiple of a base value. The second postulate was that the electrons did not radiate despite the fact that they were moving along curved paths and were being accelerated as a consequence. The fact that charged particles do radiate when accelerated was well established both from an experimental and a theoretical point of view.

In 1926, de Broglie and Schroedinger introduced independent alternate but equivalent approaches to the discipline now known as quantum mechanics. This new mechanics was not only able to account for the quantiza-

tion of electron energies in atoms and the quantization of the oscillators in black-body radiators, but included the whole of classical mechanics, which turned out to be only an approximation, but a very good one whenever it was applied to objects massive in comparison with atoms.

5.2 Wave mechanics

It is now recognized that the seeds of quantum mechanics were already present in the Hamilton-Jacobi formulation of classical mechanics. Indeed, one of the reasons that quantum mechanics seems so alien to the engineer is that the operations and semantics of quantum mechanics are borrowed from the Hamiltonian formulation, whereas the engineer is usually educated in the Newtonian formulation of dynamical problems. Despite this inherent relation between quantum and classical mechanics there is a fundamental difference over the question of what is knowable in principle. One of the basic assumptions of classical mechanics is that it is possible to measure position and velocity simultaneously. In the macroscopic realm, this seems to be self-evident, but when bodies are very small, any attempt to measure position will disturb the momentum in an unpredictable way. Similarly, any attempt to measure momentum will disturb the position unpredictably. This is known as the *uncertainty principle*, and in quantitative terms is described by the equation

$$\Delta p \cdot \Delta q \approx \frac{h}{2\pi} \tag{5.1}$$

where p is momentum, q is a position coordinate, and h is Planck's constant.

The reader might well object that it is equally improper to define quantities such as the stress and strain tensors, the current density vector, and others in terms of a limiting process when we intend to apply these concepts to matter, which is inherently discontinuous. This objection is valid, but we have recourse to empiricism here and find that we may treat matter as a continuum so long as we deal with large numbers of particles. However, the concept of a current density, for example, simply becomes an exercise in nonsense when applied to very small current flows or very small cross-sectional areas.

Since, in dealing with atomic systems, we may not formulate a mechanics which requires a simultaneous knowledge of position and momentum, upon what can the formulation be based? The answer is that we work with a probability function. We define a quantity ψ which is a function of time and a set of coordinates such that the probability of a particle's having a given set of coordinates at a particular time is $\sqrt{\psi\psi^*}$, where ψ^* is the complex conjugate of ψ. The fundamental equation relating ψ to the other system param-

eters is the Schroedinger wave equation

$$\nabla^2\psi + \frac{8\pi^2 m}{h^2}[E - V]\psi = 0 \tag{5.2}$$

where E is the total energy of the system and V is the potential energy. One of the properties of this basic equation is that energy plays a primary role, and position and momentum are accorded equal but secondary importance. The reason is that Eq. (5.2) has its antecedents in the Hamiltonian approach to classical mechanics. In contrast, Newtonian dynamics gives most of the stage to forces, leaving energy and momentum the supporting roles.

Many of the consequences of treating atomic and molecular systems in terms of quantum mechanics will seem very strange to those who have not had any previous contact with this discipline. The reader is reminded that we have been forced to adopt this approach by empirical considerations. Common sense tells us that no matter how many times we hurl an object at a closed door, the object will not penetrate the door without making a hole in it. Nevertheless, an atomic particle flung at an energy barrier with insufficient energy to surmount the barrier still has some probability of disappearing at one side of the barrier and reappearing on the other side. Furthermore, the probability of penetration can be calculated in terms of the height and thickness of the energy barrier by applying wave mechanics to the problem. The point of all of this is that we will use quantum mechanical arguments which seem bizarre because they have no counterpart in everyday sensual perception but which are based upon a theory which itself is firmly grounded upon experimental observation. The only reason these ideas may seem alien is that experimental techniques have extended our range of perception into a realm beyond the one we have been accustomed to from childhood.

One of the crucial experimental points upon which wave mechanics is based is the Davisson-Germer experiments in which it was found that electrons passing through a thin crystal are diffracted by the crystal. They show a diffraction pattern which, in both its intensity and the spacing of the maxima of the pattern, can be explained completely by assuming that the electrons are waves having a wavelength related to their momenta by the equation

$$\frac{p}{h} = \frac{1}{\lambda} \tag{5.3}$$

This does not mean that electrons are "really" waves, and in point of fact other experiments may be devised in which their corpuscular nature seems to be equally well demonstrated. The question of whether electrons are "really" waves or particles is hardly pertinent so long as we have a self-consistent comprehensive method of predicting and describing their behavior —at least in principle—under any circumstances.

Just as electrons exhibit the characteristics of waves under certain circumstances, light also exhibits some of the characteristics of particles, notably in the photoelectric effect. The light corpuscles are called *photons*. These have zero rest mass, always travel at the speed of light, and have an energy related to their frequency by

$$E = h\nu \tag{5.4}$$

These also have momentum which is related to the energy through

$$p = \frac{hc}{\lambda} \tag{5.5}$$

where c is the velocity of light.

5.3 Atomic spectra and atomic structure

The hydrogen atom is sufficiently simple that Eq. (5.2) may be applied directly and solved in closed form. This is especially rewarding, since it provides a direct explanation of the spectra of hydrogenlike atoms, gives us direct access to a deduction of the structures of more complex atoms, and provides information that is basic to a discussion of bonding. This bonanza is comparable to finding a trout stream with more fish than fisherman within an hour's drive of Manhattan Island.

The simplicity of the treatment is a result of dealing with a system containing only two particles, a proton and an electron. The relative masses of the two particles is an added advantage, since we may choose a set of coordinates with their origin coincident with the proton and thus may deal entirely in terms of the distribution of both particles about the center of mass. The potential energy of the system will be due entirely to the attractive forces between the two charged particles and will be zero when the particles are infinitely far apart. Thus

$$V = -\frac{\epsilon^2}{r}$$

and Eq. (5.2) becomes

$$\nabla^2\psi + \frac{8\pi^2 m}{h^2}\left(E + \frac{\epsilon^2}{r}\right)\psi = 0 \tag{5.6}$$

The appropriate boundary conditions are

$$\psi = 0 \text{ when } r = \infty, \qquad \int_0^\infty \sqrt{\psi\psi^*}\, dr = 1$$

The first of these requires that the probability of the particles' being infinitely far apart is zero and the second requires that the probability of the electrons being somewhere is unity.

We do not wish to go into details of the solution but merely to describe the results. The problem was formulated in spherical coordinates and we find that the solutions fall into two categories: those in which ψ is a function of r alone and those in which ψ is a function of r, θ, and ϕ.

Solutions in which ψ is a function of r alone are obtained only for specific values of the total energy

$$E = -\frac{2\pi^2 e^4 m}{h^2 n^2} \tag{5.7}$$

where n is a nonzero integer and is called the *principal quantum number*. Normally the system will reside in the lowest energy state, $n = 1$, but it may be excited to higher states by a variety of means. Its existence in a higher energy state will only be transient, and when it undergoes a transition to a lower state it will emit a photon whose frequency is related to the energy change of the transition by the equation

$$E = h\nu \tag{5.8}$$

This accounts for the general features of the spectra of hydrogenlike atoms. When viewed through a spectroscope of low resolving power, the spectrum will consist of lines whose frequencies are related through Eq. (5.8) to the transitions shown in Fig. 5.1.

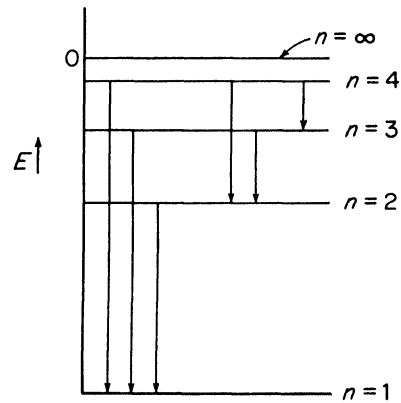


Figure 5.1 Schematic of possible transitions between different energy levels in an atom leading to spectroscopic emission lines.

For the moment the discussion has been limited to consideration of solutions in which ψ is a function of r only. Since this is the case, $\psi\psi^*$ will also be spherically symmetrical about the origin. By plotting $\sqrt{\psi\psi^*}$ versus r for various values of n, we may show how the probability of the electron's being a certain distance from the nucleus varies with the different values of n. It is possible to measure the radius of a free hydrogen atom by a number of different techniques. These all give values close to r_0 of Fig. 5.2. This is significant since in the ground state $n = 1$ and the function ψ has its maximum at r_0. In other words, r_0 is the most probable distance from the nucleus to the electron. An alternate interpretation of the curves shown in Fig. 5.2 is to consider the electron itself as being distributed in space with a density equal to $\sqrt{\psi\psi^*}$. For brevity, the electron density distribution function is called an *orbital*.

When we consider the general solution of Eq. (5.6) where ψ may also be a function of θ and ϕ, the quantization of the energy of the system is more com-

plex and requires an additional quantum number, l. This is called the *azimuthal quantum number* and it may assume values, $l = n - 1$, $l = n - 2$, . . . , 0. The various permissible energy levels of the system are determined by combinations of n and l. The principal quantum number determines the primary energy level of the system while the various values of l specify more

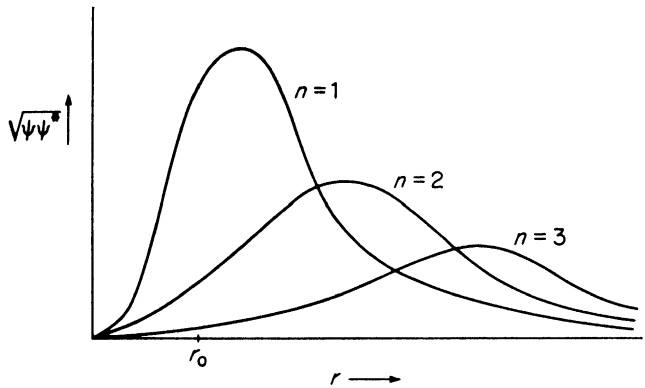


Figure 5.2 Electron density distributions around the nucleus of a hydrogen.

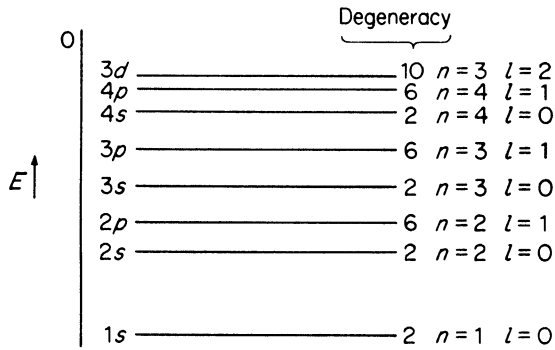


Figure 5.3 Distribution of quantum states of an isolated hydrogen atom. The energy scale is distorted.

closely spaced sublevels within each principal level. Figure 5.3 shows an energy level diagram of the hydrogen atom with the various energy levels labeled according to the combinations of quantum numbers that produce them.

For values of l greater than zero, ψ is no longer spherically symmetrical about the origin. In the cases where l equals 1, the surfaces upon which ψ is constant form a set of mutually perpendicular lemniscoids which represent the shapes of the electron orbitals. These are shown in Fig. 5.4.

Although the principal and azimuthal quantum numbers determine the energy of the system, the system may be in a number of different states having the same energy. Such states are called *degenerate* and they arise because in addition to the quantization of energy required for solutions of Eq. (5.6), it is also necessary to quantize angular momentum. The orbital angular

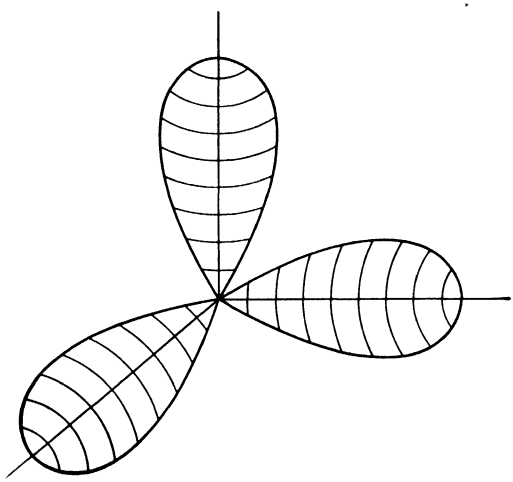


Figure 5.4 p-electron orbitals.

momentum is quantized by a set of numbers called *magnetic quantum numbers* designated by the letter m. These may have the values, $\pm l$, $\pm l - 1$, . . . , 0. Each of the different lemniscoids associated with the quantum number $l = 1$ corresponds to a different value of m. The *electron spin quantum number* is designated as s and may have the values, $\pm\frac{1}{2}$. Thus, there are two possible states of the system when n equals 1. These are the states determined by the two possible values of the spin quantum number. When n equals 2, the system may assume any of eight different states distributed among two closely spaced energy levels:

$$l = 1 \begin{cases} m = 1, & s = +\frac{1}{2}, \\ m = 1, & s = -\frac{1}{2}, \\ m = 0, & s = +\frac{1}{2}, \\ m = 0, & s = -\frac{1}{2}, \\ m = -1, & s = +\frac{1}{2}, \\ m = -1, & s = -\frac{1}{2}, \end{cases} \qquad l = 0 \begin{cases} m = 0, s = +\frac{1}{2} \\ m = 0, s = -\frac{1}{2} \end{cases}$$

For n equals 3, there are 18 states distributed among three sublevels; for n equals 4, there are 36 states in four sublevels, and so forth.

The presence of an external electric field will split the degenerate states into different energy levels. This occurs for two reasons. When l is greater

than zero, the electron densities are distributed asymmetrically, and for different values of m some will have higher energies than others because of their location relative to the electric field. In addition, the potential energy term in Eq. (5.6) will have a directional component. In a similar manner, the presence of an external magnetic field will also resolve the degenerate states into different energy levels.

Frequent reference is made to the various possible energy levels and a special notation is useful in designating them. States in which $l = 0$ are denoted s-states, those in which $l = 1$, p-states, those in which $l = 2$, d-states, and those in which $l = 3$ are called f-states. The letters are taken from the working jargon of spectroscopists. The principal quantum level is denoted by a subscript before the letter, and the number of electrons in the level is denoted by a superscript afterwards. This information may be omitted when it is not germane. Thus ${}_2s$ refers to the states having an energy level for which the principal quantum number is 2 and the azimuthal quantum number is 0. The symbol, ${}_5d^2$, states that there are two electrons in states whose energies are determined by the principal quantum number 5 and the azimuthal quantum number 2.

Technically the various states of the hydrogen atom are defined only in terms of the way the electron and the proton interact with each other. However, it is useful to regard the various states as having a virtual existence independently of the presence of the electron. We must then recognize that when an electron is added to the system, it will populate one of the states of the system, but it may also change the energies of all of the states of the system.

With the background of the previous paragraphs, we can now discuss the structure of atoms more complex than hydrogen. In quantum mechanics, many-body problems are inherently more difficult than two-body problems, just as in classical mechanics. Without attempting to be rigorous, we may still make certain plausible deductions about the structures of more complex atoms than hydrogen. Let us begin by considering a bare atomic nucleus of atomic number Z ($Z > 1$). This nucleus will have virtual states associated with it which bear a one-to-one correspondence to the states of the hydrogen atom. If a single electron is added to the system, it will be identical with the hydrogen atom, except that the principal energy levels will be computed from

$$E = -\frac{2\pi^2 e^4 m Z^2}{h^2 n^2} \tag{5.9}$$

and the electron orbitals will have their maxima closer to the nucleus.

Now if a second electron were added to the system, two factors would come into play. There is an empirical law known as the *Pauli exclusion principle*, which may be stated in the form that no two particles of a system may have an identical set of quantum numbers. Thus, when the second electron is added to the system, it will not go into the state of lowest energy

but will enter the *empty* state having the lowest energy. In this case, both $_1s$ states have equal energies, and the only restriction upon the second electron is that its spin must oppose the spin of the electron already present.

When the second electron is added to the system, mutual repulsion

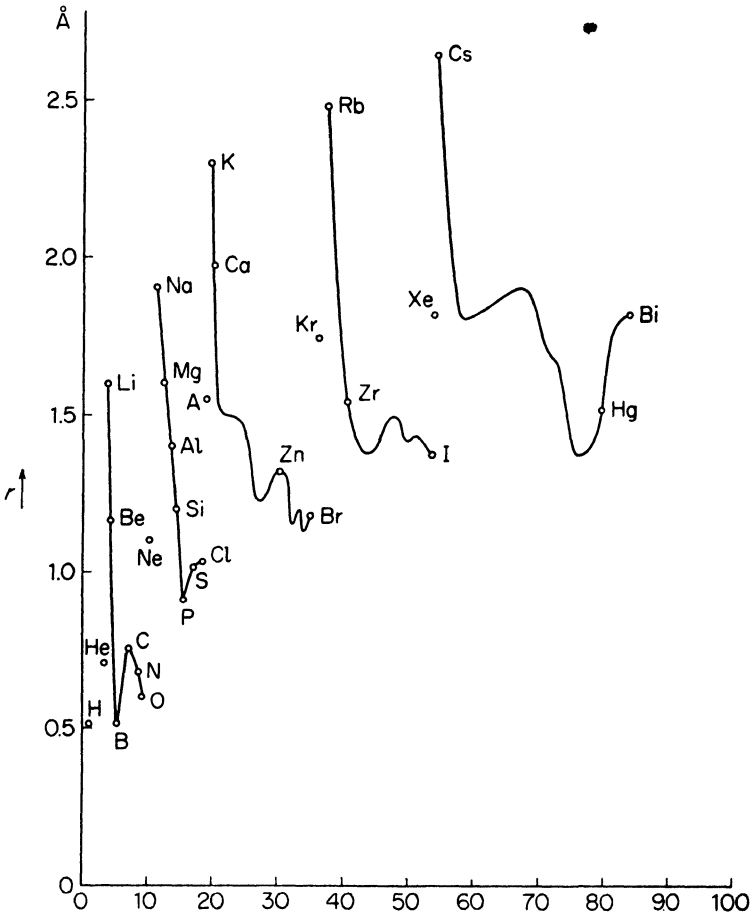


Figure 5.5 Radii of atoms vs. atomic number (Goldschmidt, *Ber. Dtsch. Ges.* **60,** 1270, (1927).)

between it and the first electron will increase the energies of all states in the system. However, it will not change the general form of their distribution in energy. If a third electron is added to the system, it will occupy a $_2s$ state, a fourth would occupy a $_2s$ state, a fifth a $_2p$ state, and so forth until the number of electrons was equal to the atomic number Z. In this manner we may deduce the electronic structure of all the lighter elements. The atomic structures of all of the elements are given in Table 5.1. It is note-

Table 5.1 *Ground state electronic configuration of the elements*

At. No.	Element	Quantum Level																	
		$1s$	$2s$	$2p$	$3s$	$3p$	$3d$	$4s$	$4p$	$4d$	$4f$	$5s$	$5p$	$5d$	$5f$	$6s$	$6p$	$6d$	$6f$	$7s$
1	H	1																		
2	He	2																		
3	Li	He Core	1																	
4	Be		2																	
5	B		2	1																
6	C		2	2																
7	N		2	3																
8	O		2	4																
9	F		2	5																
10	Ne		2	6																
11	Na	Neon Core			1															
12	Mg				2															
13	Al				2	1														
14	Si				2	2														
15	P				2	3														
16	S				2	4														
17	Cl				2	5														
18	A				2	6														
19	K	Argon Core						1												
20	Ca							2												
21	Sc						1	2												
22	Ti						2	2												
23	V						3	2												
24	Cr						5	1												
25	Mn						5	2												
26	Fe						6	2												

		$1s$	$2s$	$2p$	$3s$	$3p$	$3d$	$4s$	$4p$	$4d$	$4f$	$5s$	$5p$	$5d$	$5f$	$6s$	$6p$	$6d$	$6f$	$7s$
27	Co						7	2												
28	Ni						8	2												
29	Cu						10	1												
30	Zn						10	2												
31	Ga						10	2	1											
32	Ge						10	2	2											
33	As						10	2	3											
34	Se						10	2	4											
35	Br						10	2	5											
36	Kr						10	2	6											
37	Rb	Krypton Core										1								
38	Sr											2								
39	Y									1		2								
40	Zr									2		2								
41	Cb									4		1								
42	Mo									5		1								
43	Tc									5		2								
44	Ru									7		1								
45	Rh									8		1								
46	Pd									10										
47	Ag									10		1								
48	Cd									10		2								
49	In									10		2	1							
50	Sn									10		2	2							
51	Sb									10		2	3							
52	Te									10		2	4							
53	I									10		2	5							
54	Xe									10		2	6							
55	Cs	Xenon Core														1				
56	Ba															2				
57	La													1		2				

Table 5.1 *Ground state electronic configuration of the elements (Continued)*

At. No.	Element	Quantum Level																	
		$1s$	$2s$	$2p$	$3s$	$3p$	$3d$	$4s$	$4p$	$4d$	$4f$	$5s$	$5p$	$5d$	$5f$	$6s$	$6p$	$6d$	$7s$
58	Ce										1			1		2			
59	Pr										2			1		2			
60	Nd										4					2			
61	Pm										5					2			
62	Sm										6					2			
63	Eu										7					2			
64	Gd										7			1		2			
65	Tb										8			1		2			
66	Dy										10					2			
67	Ho										11					2			
68	Er										12					2			
69	Tm										13					2			
70	Yb										14					2			
71	Lu										14			1		2			
72	Hf										14			2		2			
73	Ta										14			3		2			
74	W										14			4		2			
75	Re										14			5		2			
76	Os										14			6		2			
77	Ir										14			9					
78	Pt										14			9		1			
79	Au										14			10		1			
80	Hg										14			10		2			
81	Tl										14			10		2	1		
82	Pb										14			10		2	2		

At. No.	Element	Quantum Level																	
		$1s$	$2s$	$2p$	$3s$	$3p$	$3d$	$4s$	$4p$	$4d$	$4f$	$5s$	$5p$	$5d$	$5f$	$6s$	$6p$	$6d$	$6f$	$7s$
83	Bi										14			10		2	3			
84	Po										14			10		2	4			
85	At										14			10		2	5			
86	Rn										14			10		2	6			
87	Fr	Radon Core																		1
88	Ra																			2
89	Ac																	1		2
90	Th																	2		2
91	Pa														2			1		2
92	U														3			1		2
93	Np														5					2
94	Pu														6					2
95	Am														7					2
96	Cm														7			1		2

worthy that the $_4s$ states actually have a lower energy than the $_3d$ states and the $_5s$ and $_5p$ states are lower in energy than the $_4f$ states.

Reference to Table 5.1 shows a certain amount of regularity in the way additional electrons populate the various quantum levels as the atomic number increases. This produces a periodicity in both the physical and the chemical properties of the elements. For example, the periodic variation in atomic radius is shown in Fig. 5.5. A periodic table of the elements is shown on page 279.

5.4 The covalent bond

Ultimately any treatment of bonding must start from the wave equation. Since the hydrogen molecule is the simplest of all possible molecules, it was

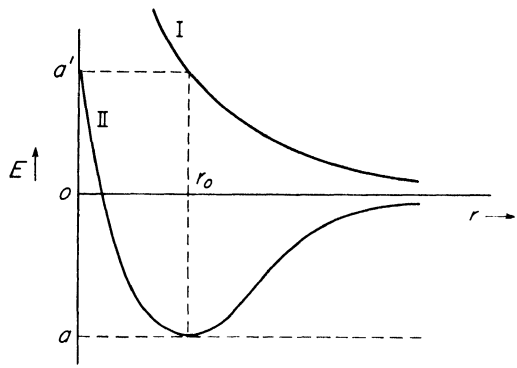


Figure 5.6 Energy as a function of interatomic distance for a typical covalent bond in a diatomic molecule.

the first to yield to a complete quantum mechanical treatment. This is a four-body problem (two protons, two electrons) and the wave equation could be solved only by numerical methods. This was done prior to the advent of electronic computers; the achievement was Herculean purely from the standpoint of the burden of the arithmetic involved.

As a result of this treatment of the hydrogen molecule, it was found that as two normal hydrogen atoms approached one another, the total energy of the system could vary with the distance between atoms in either of two ways as shown in Fig. 5.6. In curve I, the forces between the atoms are everywhere repulsive and no bonding occurs. In curve II, however, there is an energy minimum at an interatomic distance r_0 which corresponds to a bond's being formed between the two atoms.

For each hydrogen atom in the system there are two $_1s$ states; consequently the hydrogen molecule will contain four $_1s$ states and two electrons. When the two atoms are bonded, the energies of the states are disturbed in such a

way that the energies of half the states are diminished, while those of the other half are augmented. In Fig. 5.6 two states will have the energy of point a and the other two will have the energy of point a'. The two of lowest energy will normally be occupied by the bonding electrons, while the two higher-energy states will normally be empty. The bonding electrons will always have opposed spins.

This splitting of states is a general characteristic of the formation of bonds and will be an important point in the development of the band theory in a later section. In the case of the hydrogen molecule, all of the energy states split in this manner but only the $_1s$ state is of interest.

Although the wave equation itself is the only valid point of departure for a quantitative discussion of the covalent bond, in a qualitative sense we may think of the covalent bond as being formed from an overlapping of electron orbitals in which the spins of the electrons are opposed. The electron density in a bond is roughly equal to the sum of the electron densities of the overlapping orbitals. The orbitals of electrons in states other than s-s bonds are oriented in specific directions with respect to one another, and these directions are closely related to the observed bond angles in multivalent atoms. Frequently orbitals in which the electron spins are not opposed are referred to as *antibonding* orbitals.

5.5 Resonance and exchange energy

The preceding treatment of the covalent bond could be checked experimentally on two points, the bond energy D and the interatomic distance r_0. In the initial treatment, the correct value of r_0 was obtained but the bond energy was computed to be about 66 kcal/mole rather than the observed 103 kcal/mole. Later this discrepancy was explained by the quantum mechanical phenomenon known as *resonance*.

If a system can exist in two or more states of nearly the same energy, it will not remain in any one but will "resonate" between all of them. This shifting from one state to another occurs very rapidly, and has the net effect of creating a hybrid state whose energy is lower than that of the individual states. The characteristics of a hybrid state will be related to but different from those of the resonating states from which it is formed. This is a basic phenomenon which will enter into many of the arguments which we shall develop later.

As a very simple illustration of resonance, we might consider a particle in a potential well shaped as shown in Fig. 5.7. The states A and B are identical in energy, and in classical mechanics we would consider that either state would be equally probable and that the system would remain permanently in either one or the other. However, in quantum mechanics, if the potential hump in the center is not too high or broad, the system will

resonate between states A and B to form a hybrid state in which the wave function of the particle would have the shape of Fig. 5.8. Physically, this

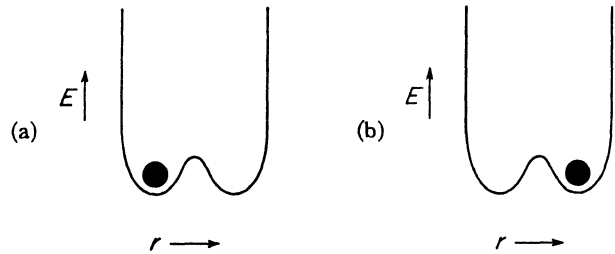


Figure 5.7 Schematic of a condition leading to resonance.

means that the particle might be on side A or on side B, but it also has a nonzero probability of being in between. In brief, it is in a state which is actually a hybrid of A and B. Such a hybrid state invariably has a lower energy than either A or B.

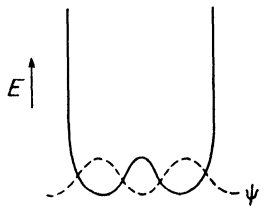


Figure 5.8 Dashed line represented the wave function of a particle in a hybrid state.

This decrease in energy is usually termed *exchange energy* when an exchange of configurations is involved. However, the concept is broader than this and a system may resonate between states characterized by different quantum numbers just as well without involving any geometrical considerations.

In the case of the hydrogen molecule, about 40% of the bond energy is due to resonance. If we label one of the hydrogen atoms 1 and the other 2, the possible bonding states of the system would be:

1. both electrons in $_1s$ state of nucleus 1, spins opposed
2. both electrons in $_1s$ state of nucleus 2, spins opposed
3. one electron in $_1s$ state of nucleus 1, spin $+\frac{1}{2}$, and the other electron in $_1s$ state of nucleus 2, spin $-\frac{1}{2}$
4. same as (3) with the spins reversed.

Actually the hydrogen molecule goes into a hybrid state and has an energy lower than any of these.

5.6 Hybridization of bonds

According to the periodic table of the elements on p. 279, we find that hydrogen has a valence of one, helium a valence of zero, lithium a valence of one, beryllium a valence of two, and boron a valence of three. These are the

first five elements listed in Table 5.1, and it will be interesting to correlate their observed chemical valencies with their electronic structure. Hydrogen behaves according to expectation; it has only one electron and could hardly have a valence other than one. Helium has both $_1s$ states full, and these electrons would not be available for bond formation. Lithium has one electron in a $_2s$ state and this single electron gives it a valence of one. Beryllium has both the $_1s$ and $_2s$ states filled and we might well expect it to have a zero valence. Using the same line of reasoning, we again expect that only the $_2p$ electron of boron would be available for bond formation and that it would accordingly have a valence of one. Actually the $_2s$ and $_2p$ states are very nearly of the same energy. Because of this, the electrons in the second principal quantum levels of beryllium, boron, and carbon resonate between *s* and *p* states to form hybrid *s-p* states.

It will be recalled that *s*-orbitals are not directed in space although *p*-orbitals are. Normally bonds formed by the overlapping of *p*-orbitals would be mutually perpendicular. Hybrid *s-p* orbitals are also directed in space in a manner determined by the number of *p*-orbitals involved in the hybrid. The orbitals formed from the hybridization of an *s*-orbital with one *p*-orbital will be directed at 180° to one another. The hybrids formed from one *s*- and two *p*-orbitals are coplanar and are directed at 120° with respect to one another. Those formed from three *p*- and one *s*-orbitals are directed tetrahedrally.

5.7 Elementary band theory—bonding in metallic and covalent crystals

We must distinguish between those crystalline solids in which individual molecules are bonded together by van der Waals forces to form a crystal and those solids in which every atom in the solid is bonded to its neighbors by covalent bonds. The former are called *molecular solids*, while the latter are called *covalent solids* and it is these which are considered in this section along with metals.

The metallic bond and the bonding in covalent crystals, while quite different in their general character, are both dependent upon the density and distribution of quantum states in the solid and the way these are populated by electrons. This electronic structure of solids may be described by either of two equivalent theories. One of these is called the *band theory* and the other the *zone theory*. The former is under discussion in this section and the latter is considered in the next section. Both are closely linked with the bonding in the solid, and both are important in explaining the electrical properties of solids. Here the major objective is to show how the band theory provides a unifying base through which the metallic bond can be related to the covalent

bond. However, ultimately this whole discussion of bonding is directed toward explanations of various properties, and the band theory as well as the zone theory will form the basis of many explanations in later chapters.

As a point of departure in formulating the band theory, we must go back to the formation of the covalent bond as discussed in Sec. 5.4. It will be recalled that when a covalent bond is formed, the states of the energy levels containing the bonding electrons split into two equal groups, one of which is promoted in energy and one of which is depressed in energy. The behavior of carbon in forming the diamond is fairly typical and will serve as an example. The diamond structure is shown in Fig. 5.9. Each atom of carbon forms four hybrid *s-p* bonds. These are directed at tetrahedral angles linking each atom in the crystal with its four nearest neighbors. The entire crystal is a single molecular system. Each carbon atom contributes eight states to the system and four electrons. The states split into a set of high-energy vacant states and low-energy bonding states, each band containing four times as many states as there are atoms.

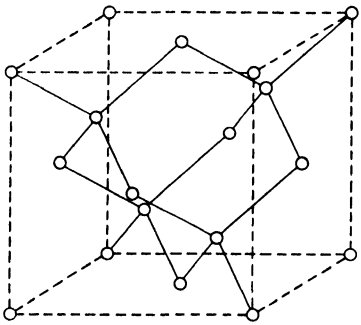


Figure 5.9 The diamond lattice.

Since the entire crystal is a single system, these states are no longer descriptive of individual atoms, but characterize the electronic structure of the crystal

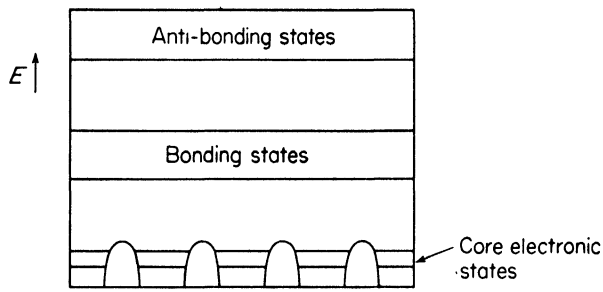


Figure 5.10 Band structure in a covalent solid.

as a whole. Three of the four quantum numbers lose their significance (spin is the exception), and the application of the exclusion principle requires that both the bonding states and the antibonding states be distributed over a range of energies rather than constituting a degenerate set all having the same energy. However, the energies of the states will be closely spaced and they will form a pair of bands as shown schematically in Fig. 5.10.

The upper band is called the *conduction band* and the lower is called the

valence band. Physically most of the electrons of the system will be in one of the valence-band energy states. These are actually located in carbon-carbon bonds. Nevertheless, they are free to move about the crystal but are coupled in pairs having equal but opposite momenta. Thus, even though the electrons in the valence band are mobile, they cannot conduct electricity if the band is filled. There are, however, a number of mechanisms by which the electrons may be excited out of the valence band into the conduction band where they are capable of carrying an electric current. This topic must be deferred to a later chapter, since our major concern here is the bonding characteristics of covalent solids.

A large number of different types of solid-state electronic devices are based upon single covalent crystals. The use of germanium and silicon in diodes and transistors is well known, but there are a host of other devices such as Hall amplifiers, photoconductive switching devices, thermoelectric devices, and the like which are also based upon crystals of this class. These crystals are not limited to the Group IV elements. Compounds of the Group III and Group V elements such as indium antimonide, boron nitride, gallium arsenid, etc., also fall into this class and even have the diamond structure. Even when both constituents of the compound are metals, the compounds themselves are covalent. Other compounds such as Bi_2Te_3 and Sb_2Se_3 are also covalent crystals but do not have the diamond structure.

Even though the metallic bond is considered to be quite different from the covalent bond, the differences are really quite superficial. When a group of atoms are condensed into a metallic crystal, again the states in the highest quantum level which is populated by electrons in the individual atoms split into bands as described previously. If the outer electrons are in the second principal level, each atom contributes eight states to the assembly which split into two groups of four. When the principal quantum level of the outer electrons is three, the *d*-states differ so much in energy from the *s*- and *p*-states that they may be ignored, and again each atom contributes eight states which split into two bands. When the outer electrons are in the fourth principal level, the *s*-, *p*-, *d*-, and $_3d$-states will all enter into forming bands.

An assemblage of atoms will be metallic in its properties either when there are fewer electrons per atom than there are states per atom in the highest energy band containing electrons or when the upper and lower bands are spread over such a broad range of energies that they overlap leaving no gap between the bands. These situations are illustrated in Figs. 5.11 and 5.12. On the basis of these ideas we might expect all of the elements in Groups I, II, and III of the periodic table to be metals. In the second case, we might expect all elements having a relatively high atomic number to be metals. Although there are exceptions, these generalizations hold fairly well.

Elemental metals generally crystallize in a form such that each atom has eight or twelve nearest neighbors. Usually there will not be enough electrons

to form permanent bonds between each atom and all of its neighbors. Instead, transient bonds resonate between different neighbors. The large numbers of alternate configurations having the same energy provide sufficient resonance energy to give the metallic bond a high degree of strength even though the individual bonds are only transient in character. This transient character also gives metals a high degree of malleability since the bonds are not directed in space. In contrast, the bonds in a covalent crystal

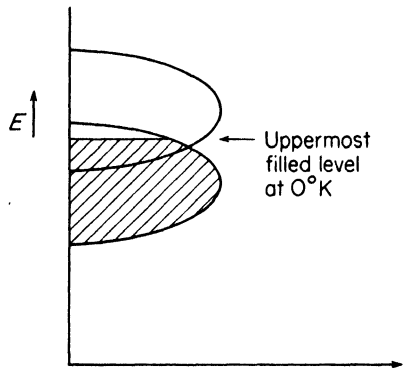


Figure 5.11 Density of states distribution showing overlapping bands.

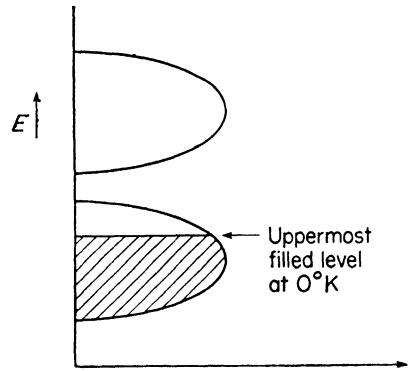


Figure 5.12 Density of states distribution showing a forbidden region between bands.

are directed in specific directions, and if strained beyond a certain point they will lose their strength. Thus covalent crystals are typically hard and brittle.

Our initial criterion of a metal demanded that the bonding electrons reside in an incomplete band of energy levels. The electrons can be excited into the empty states by a very small amount of energy; when they are so excited, they become decoupled, so that not all of the electrons will be paired with other electrons having opposite momenta. Thus the electrons may carry an electric current.

5.8 Zone theory

The band theory as developed in the previous section, is a logical extension of what is known of the nature of the covalent bond. The general outline of the band structure deduced in this manner is correct, but aside from knowing that the energy states split and form bands, we have no quantitative information about how the states are distributed in energy in the bands. When a large number of atoms are condensed into a crystalline system, there is an alternate method of approaching the problem which will provide this information as well as some other quantitative information about the system. This approach is known as the *zone theory*.

Our development of the zone theory will be similar to the procedure of deducing the distribution in energy of the quantum states of a free atom. We shall first make an assumption about the potential distribution inside the solid and solve the wave equation for the case where only one electron is present in the solid. This will give us the distribution of states in energy. We will then assume that these states may be populated by additional electrons without changing the number of states or their distribution in energy.

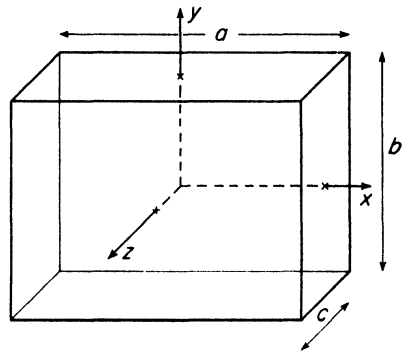


Figure 5.13 Coordinate system and boundaries used in the development of the zone theory.

We may simplify the boundary conditions governing the solutions of the wave equation with no loss in generality by specifying that the solid be rectangular in shape. We assume that the potential energy is everywhere zero inside the solid, and jumps abruptly to infinity at the surfaces of the solid. We choose a coordinate system with the origin located in the center of the solid and with the x axis parallel to the side of length a, the y axis parallel to the side of length b, and the z axis parallel to the side of length c. Since the potential energy is zero everywhere inside the body, the wave equation assumes the form

$$\nabla^2\psi + \frac{8\pi^2 mE\psi}{h^2} = 0 \tag{5.10}$$

with the boundary conditions

$$\psi = 0, \quad x = \pm\frac{a}{2}, \quad y = \pm\frac{b}{2}, \quad z = \pm\frac{c}{2}$$

Solutions of the form

$$\psi = \psi_0 \cos n_1\pi\frac{x}{a}\cos n_2\pi\frac{y}{b}\cos n_3\pi\frac{z}{c} \tag{5.11}$$

where the n's are odd integers will fit these boundary conditions provided that the energy is restricted to the values

$$E = \frac{h^2}{2m}\left[\left(\frac{n_1}{2a}\right)^2 + \left(\frac{n_2}{2b}\right)^2 + \left(\frac{n_3}{2c}\right)^2\right] \tag{5.12}$$

These, however, are not the only solutions possible. The solution

$$\psi = \psi_0 \sin n_1\pi\frac{x}{a}\sin n_2\pi\frac{y}{b}\sin n_3\pi\frac{z}{c} \tag{5.13}$$

also fits the boundary conditions provided that the n's are even integers. Again these are valid only when the energy is restricted to the values of Eq. (5.12).

Since $V = 0$, E is the kinetic energy, $mv^2/2$, and substituting $p = mv$, Eq. (5.12) reduces to

$$\frac{p^2}{h^2} = \left(\frac{n_1}{2a}\right)^2 + \left(\frac{n_2}{2b}\right)^2 + \left(\frac{n_3}{2c}\right)^2 \tag{5.14}$$

The momentum of a particle is related to the wavelength of the associated wave function through $p/h = 1/\lambda$, and Eq. (5.12) specifies the restrictions placed upon the wavelength of the particle by the external dimensions of the solid. Each of these wavelengths specifies a particular quantum state, and these states have the energies specified by Eq. (5.12).

Because of the de Broglie relation $p/h = 1/\lambda$, Eq. (5.14) may be written in the form

$$\frac{1}{\lambda^2} = \left(\frac{n_1}{2a}\right)^2 + \left(\frac{n_2}{2b}\right)^2 + \left(\frac{n_3}{2c}\right)^2 \tag{5.15}$$

This is identical in form with the dot product of a vector with itself, suggesting that it could be written in the form

$$\mathbf{W} = \frac{\mathbf{n}_1}{2a} + \frac{\mathbf{n}_2}{2b} + \frac{\mathbf{n}_3}{2c}$$

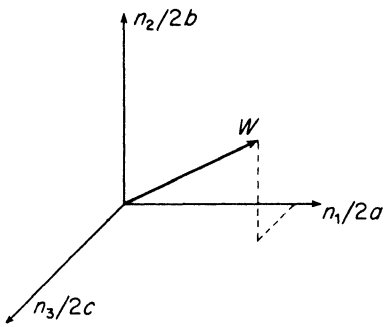


Figure 5.14 Wave number vector in n-space.

where $\mathbf{n}_1/2a$, $\mathbf{n}_2/2b$, and $\mathbf{n}_3/2c$ are the magnitudes of a set of mutually orthogonal components of a vector $\mathbf{W}$.[1] As shown in Fig. 5.14, the terminus of the vector $\mathbf{W}$ is a unique point depending upon the values given the $\mathbf{n}$'s. The totality of all points determined in this way constitutes a space called variously *phase space*, *momentum space*, or in this text, *n-space*. This is a three-dimensional space whose dimensions are reciprocal lengths. We may represent increasing values of n_1, n_2, and n_3 as increasing distances along a set of Cartesian coordinates, and such a set of coordinates may be used to represent points, lines, and figures of n-space in ordinary space.

The whole purpose of inventing this n-space is to provide a convenient tool for finding the distribution of quantum states in energy. Each point in n-space will reside in a cell of volume $1/2a \cdot 2b \cdot 2c$ which contains no other point representing a permitted quantum state corresponding to a permitted reciprocal wavelength. A spherical octant of radius $1/\lambda$ in n-space will con-

[1] Most treatments of this problem use a wave number vector $\mathbf{k}$ defined by $\mathbf{k} = 2\pi(1/\lambda)$.

tain approximately

$$\frac{\frac{1}{8} \cdot \frac{4}{3} \pi \left(\frac{1}{\lambda}\right)^3}{\frac{1}{8abc}}$$

points of wavelength less than or equal to $1/\lambda$. The approximation is very good for $(1/\lambda)^3$ greater than $1/8abc$. The spherical octant was used instead of an entire sphere because negative n's represent a reversal of one component of momentum. The number of states having a reciprocal wavelength less than $1/\lambda$ per unit volume in the material under consideration then will be

$$N\,(1/\lambda) = \frac{4\pi}{3}\,(1/\lambda)^3$$

and the spectral distribution will be

$$dN\,(1/\lambda) = 4\pi\,(1/\lambda)^2\,d\,(1/\lambda) \tag{5.16}$$

In this case, we are not particularly interested in the spectral distribution *per se*, but rather in the distribution of quantum states in energy. The number of states having an energy between E and $E + dE$ may be computed from the spectral distribution in the following way. Equation (5.16) may be written

$$dN\,(1/\lambda) = \frac{4\pi p^2}{h}\,dp$$

but $p^2 = 2E/m$ and $dp = \sqrt{2/mE}$. Therefore

$$dN\,(1/\lambda) = \frac{2\pi}{h^3}\sqrt{8m^3E}\,dE \tag{5.17}$$

These quantum states each have a twofold degeneracy. That is, each is capable of being occupied by two electrons having opposed spins and moments.

Now let us criticize and refine this treatment of the solid. The most important differences between the treatment we have followed and what is known of the true situation inside the solid are:

1. Real solids do not provide an infinite potential barrier to electrons at their surfaces. For this reason the wave function will not go to zero at the physical boundary of the solid. It will have a small amplitude at the surface and will change its form to that of a standing wave whose amplitude decreases exponentially with distance from the surface. We shall consider the practical effect of this upon the distribution of states in Chap. 10. For the present we shall not try to improve the theory in this respect.

2. Since the atomic nuclei are positively charged and are relatively fixed in their positions in the lattice, they will cause a periodic variation in the

electrical potential in the solid. This is in contrast with the assumption that the potential is zero and constant.

In our refinement we shall not attempt to express the variation of potential due to the nuclei in analytical form and incorporate it into the wave equation. Instead we shall take advantage of the wave character of electrons, and utilize the fact that the lattice will cause diffraction of the electron wave when it has a wavelength equal to twice the lattice spacing in the direction of its momentum vector. In essence, the solution of the wave equation imposes a

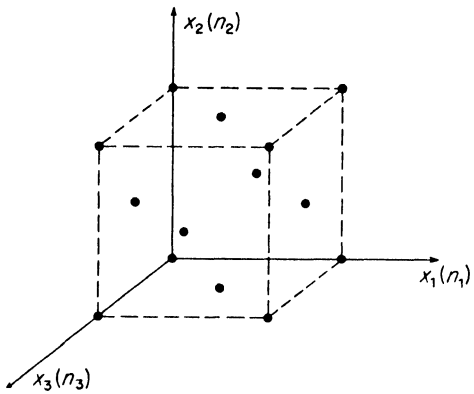


Figure 5.15 Superimposed Cartesian coordinates for n-space and x-space with atoms of a fcc basic cell located in the x-space system.

further restriction upon the wavelengths of the electrons at the long-wave end of the spectrum. The net result of these considerations is that Eq. (5.17) will govern the distribution of quantum states in energy except in regions where this distribution is disrupted by diffraction of the electron waves by the lattice.

Those points in n-space which correspond to values of $1/\lambda$ for which the lattice diffracts the electrons waves fall upon a set of surfaces which are called *Brillouin zones*. Their shape and significance can best be described by considering an example in detail. We shall choose the first Brillouin zone of the fcc lattice as our example. We choose a set of coordinate axes coincident with the crystallographic axes, and let the distance d represent the length of the edge of the basic cell. Figure 5.15 shows the basic cell of the lattice in this coordinate system. We shall also require a coordinate system describing n-space, since the Brillouin zone is a polyhedron in n-space. We may choose the axes, n_i, of this system to be coincident with the x_i system of Fig. 5.15. The spacing of the (100) planes is $d/2$, hence an electron traveling through the lattice in the x_1 direction would be diffracted by the lattice when $1/\lambda = d$.

If the electron were traveling in a direction making angles α, β, and γ

respectively with the x_1, x_2, and x_3 directions, and if we still consider only the effect of the (100) planes upon the minimum wavelength, we would have diffraction when $\lambda = d \cos \alpha$ or $1/\lambda = 1/d \cos \alpha$. This equation represents a family of wave number vectors terminating upon a pair of planes perpendicular to the n_1 axis and located a distance $1/d$ on each side of the origin. These are called *Brillouin planes.* With every set of lattice planes in x-space, there is a corresponding pair of associated Brillouin planes in n-space. The

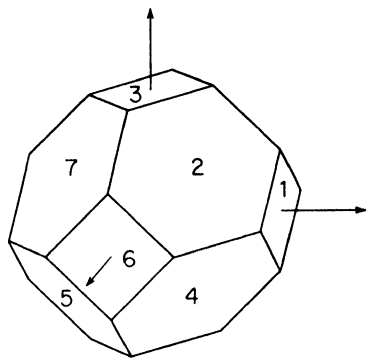


Figure 5.16 The Brillouin planes that are visible in the perspective in which the drawing is shown are numbered. These are listed below in conjunction with their distance from the origin and their associated lattice planes. The edge of the basic cell has the dimension d.

	Associated Lattice Plane	Perpendicular Distance to Origin
1.	(100)	1/d
2.	(111)	$\sqrt{3}$/2d
3.	(010)	1/d
4.	(111)	$\sqrt{3}$/2d
5.	(111)	$\sqrt{3}$/2d
6.	(001)	1/d
7.	(111)	$\sqrt{3}$/2d

Brillouin planes will be oriented in n-space in the same way as their associated lattice planes are oriented in x-space. The Brillouin planes will be located in n-space so that the perpendicular distance from them to the origin of the n_i coordinate system is equal to the reciprocal of twice the distance between the corresponding lattice planes.

If we use this procedure to construct the Brillouin planes corresponding to the (100), (010), (111), (001), $(1\bar{1}1)$, $(11\bar{1})$, and $(\bar{1}11)$ planes for the fcc lattice, we find that they intersect to form the body shown in Fig. 5.16. The region enclosed by this body is called the *first Brillouin zone.* Brillouin planes associated with other lattice planes will not intersect this body. They will, however, form bodies of larger dimensions which are called the *second, third,* etc., Brillouin zones.

Thus far we have assumed that an electron wave traveling through the lattice would be transmitted without any reflected component except for those values of momentum at which Bragg reflection would occur. Actually as

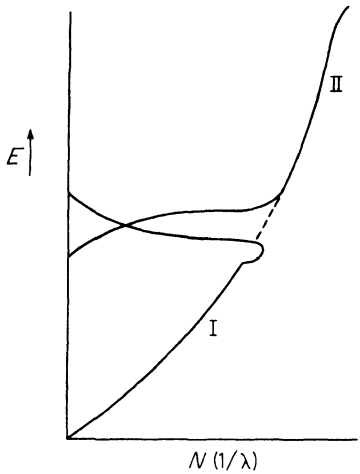


Figure 5.17 Density of states function for overlapping bands; the dashed portion of the curve represents the function if there were no lattice diffraction.

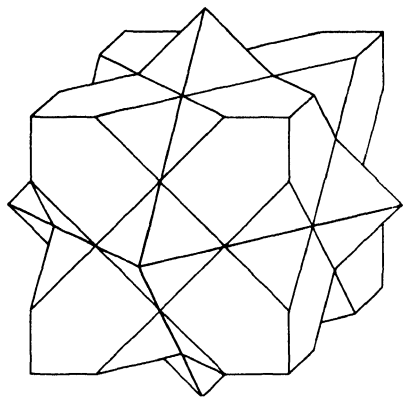


Figure 5.18 Second Brillouin zone of a fcc lattice.

the momentum approaches a zone boundary, progressively larger fractions of the electron wave are reflected. This causes the spectral distribution curve to depart from the parabolic form of Eq. (5.16) and follow the solid portion of the curve shown in Fig. 5.17. There may or may not be an overlap between the first and second zones.

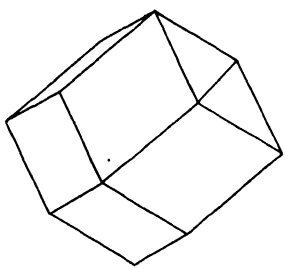


Figure 5.19 First Brillouin zone of a bcc lattice. The second zone is similar to the first zone of a fcc lattice.

To summarize, Eq. (5.17) will give a valid representation of the distribution of quantum states in energy for all states whose reciprocal wavelengths when plotted in n-space will lie inside a sphere tangent to the inner surface of the first Brillouin zone. Between this and the zone surface, the distribution of permitted wavelengths departs from the distribution expressed in Eq. (5.17); and at the zone surface, states associated with momenta of the same magnitude may be forbidden in some directions but not in others.

It was remarked earlier that the external dimensions of the body impose a low-frequency limit upon permitted values of $1/\lambda$ of electrons. When the

wavelength has the proper value for Bragg reflection, the lattice interrupts the spectrum, but electron wavelengths shorter than twice the spacing of the lattice planes in the direction of propagation are permitted. Except as these are influenced by the zone boundaries, they will also be distributed in energy

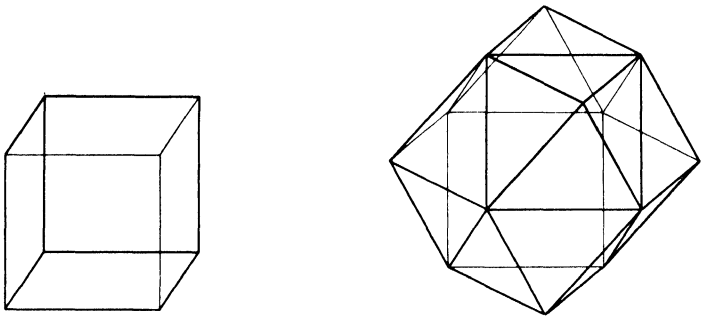


Figure 5.20 First and second Brillouin zones of a simple cubic lattice.

according to Eq. (5.17). The point in n-space representing reciprocal wavelengths of these electrons will be bounded at the low-frequency end by the first Brillouin zone boundary; those at the high-frequency end will be bounded by the second Brillouin zone boundary. The net result is a distribution of quantum states in energy as shown in curve II of Fig. 5.17.

Figures 5.18 through 5.22 show the first and second Brillouin zones of the common crystal types. The actual computation of the volume of a zone in n-space is very difficult and not a very interesting task. The volume of the first zone for the bcc structure is $2/d^3$; that of the second zone is $4/d^3$. Since each cell in n-space has a volume of $1/8abc$, the total number of cells inside the first zone will be $(2/d^3)/(1/8abc)$. We count only those cells in the positive eighth of the quadrant, consequently the number of cells per unit volume will be

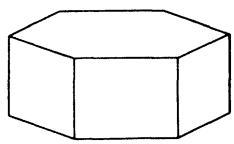


Figure 5.21 First Brillouin zone of hcp structure.

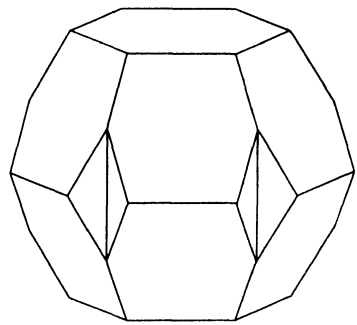


Figure 5.22 Second Brillouin zone of an hcp structure.

$$\frac{1}{8} \cdot \frac{1}{abc} \cdot \frac{2/d^3}{1/8abc}$$

In other words, the total number of states per unit volume is identical with the volume of the Brillouin zone. This is general and may be applied to any structure.

In the interior of the first Brillouin surface the quantum states have a twofold degeneracy; consequently, the number of electrons per unit volume that may be accommodated by states inside the first zone will be equal to twice the volume of that zone. Although this quantity may vary from one structure to another, the number of electrons per atom that may be accommodated by states inside the first Brillouin zone will always be two. This fact may be illustrated by considering the bcc structure. The number of states per unit volume is $2/d^3$, the number of electrons per unit volume for which there are available states is $4/d^3$, and the number of atoms per unit volume can be deduced to be $2/d^3$ by inspection of the basic cell of the structure. Consequently, two electrons per atom can be accommodated with the first zone. The volume of the first zone in n-space is always equal to the number of atoms per unit volume in Euclidean space, and the volume between consecutive zones will equal the number of atoms per unit volume of n-space.

Although the states of the first Brillouin zone have a twofold degeneracy, those of the second zone have a sixfold degeneracy, those of the third zone have a tenfold degeneracy, etc. The volume of successive zones will be an integral multiple of the volume of the first zone, where the integer is the zone number. Still using the bcc structure as an example, the volume of the second Brillouin zone is $4/d^3$, that of the third is $6/d^3$, etc. When we speak of the states in the second zone, we mean those inside the second zone but outside the first.

5.9 Applications of zone theory to metals

We have remarked earlier that the nature of the twofold degeneracy of the first Brillouin zone is such that the two electrons in each energy level have equal and opposite momenta. With the momenta coupled in such a manner, they would not be able to conduct electricity if the zone were filled. However, if a zone is only partly filled, as would be the case in monovalent metals, very small increments of energy, supplied by either thermal excitation or an electric field, would promote those electrons near the top of the band of filled levels to higher levels. Those energy levels occupied by just one electron would not exhibit this coupling of momenta and, consequently, the electrons would be free to conduct electricity. Materials having two electrons per atom thus could be insulators since the first zone would be completely filled. However, in divalent metals there is a certain amount of overlap in the distribution of quantum states in energy for the first zones. In this region of overlap an electron may go from a state in a lower zone to a state in an upper zone with only a change in the components of its momentum and no change in energy. Once it is in a state belonging to the second zone, it may be

excited to the adjacent vacant states and participate in the conduction of electricity. Even so, the electrons available for conduction in this manner will be of lower concentration than those available for conduction when the zone is half full. As a result of this, mono- and trivalent metals will generally be much better conductors than divalent metals. This generalization does not apply to the transition metals.

Another rather striking application of zone theory is in the structure of alloys. When metals of different valencies are alloyed, it is possible to have fractional values of the ratio of electrons to atoms. The value of the ratio will often be the governing factor in determining the structure of solid solutions. The relationship between the electron-to-atom ratio and structure is part of a set of rules called the *Hume-Rothery rules*, which originally were empirical. When the electron-atom ratio is less than 1.4, the fcc structure is favored. When the ratio reaches 1.4, the fcc structure becomes unstable with respect to either the bcc or the β-manganese structure. These in turn are unstable with respect to the γ-brass structure at an electron-atom ratio of 1.64. Between 1.64 and 1.75 the hcp structure is most likely to be preferred.

As early as 1926 Hume-Rothery noticed that at an electron-to-atom ratio of 1.4, the states in the first Brillouin zone of the fcc were just filled to a point where all the points in n-space representing occupied states formed a sphere tangent to the zone surface. The situation was the same for a bcc structure at a ratio of 1.5, for the γ-brass structure at a ratio of 1.64, and for the hcp structure at a ratio of 1.75. In each of these cases, the same factors are responsible for the instability of the structure after the electron-atom ratio passes a certain point. After the zone has been filled to the point where the surfaces of equal energy become tangent to the zone surface, the density of states falls off rapidly with increasing energy. Thus, a few additional electrons would raise the energy of the system much more rapidly than before. This might make it unstable with respect to a structure having a Brillouin zone that could accommodate spheres of larger radii. With the fcc structure this occurs when the sphere becomes tangent to the Brillouin plane corresponding to the (111) lattice planes. The radius of this sphere in n-space is $\sqrt{\frac{3}{2}}\,d$ and the volume of the sphere is

$$\frac{4\pi}{3} \cdot \frac{3^{3/2}}{8d^3}$$

Since a volume of $2 \cdot d^3$ corresponds to a ratio of one electron per atom, then the boundary is reached at an electron-atom ratio of $\pi\sqrt{\frac{3}{4}}$, or 1.36.

5.10 Correlation between zone theory and band theory

The zone theory, as developed in the previous section, and the band theory presented in Sec. 5.7 are different but essentially equivalent methods of describing the electronic structure of solids. The band theory evolves

directly from bonding considerations, whereas bonding is not directly involved in zone theory. Instead, the interactions of the lattice with the electrons play the dominant role in the structure of the theory. All factors considered, the latter is most useful in accounting for the properties of metals, while the former is perhaps more useful in discussing either insulators or semiconductors. However, arguments based upon either theory are equally valid, and may be chosen for their convenience or clarity in any specific application. Both of these have some correlation with the quantum states of the atoms involved, but the principal quantum number does not appear explicitly in either theory. Instead, it is understood that the energy states present in the bands of states encountered in either theory were all derived originally from the principal quantum level of the individual atoms that contained the valence electrons.

For example, in the case of crystalline silicon, the valence electrons are all in the third principal quantum level. Each atom in the structure has a total of eighteen states in this level. The ${}_3d$ level is of considerably higher energy than either the ${}_3s$ or ${}_3p$ states (see Fig. 5.3) and may be omitted from consideration. The silicon atoms are bonded in the crystal by four hybrid s-p bonds directed at tetrahedral angles. In these circumstances, each silicon atom contributes eight states to the array and four electrons. When the atoms are condensed into a solid, the states split into two bands each containing four states per atom. The lower band will normally be filled by the bonding electrons, while the upper band remains empty.

Now let us treat this same example in terms of the zone theory. Crystalline silicon has the diamond structure (see Fig. 5.7) and the atoms all lie on crystallographic planes corresponding to the Brillouin planes from which the first zone of the fcc lattice is constructed. The first has the volume of $2/d^3$, and the volume between the first and second zones is $3/d^3$. The diamond lattice contains a total of eight atoms in the basic cell; consequently, there are $8/d^3$ atoms per unit volume. The states in the first zone have a twofold degeneracy and those between the first and second zone boundaries have a sixfold degeneracy. Therefore, the first two zones can accommodate $2.4/d^3 + 6.4/d^3 = 32/d^3$ electrons per unit volume or four electrons per atom. In other words, the electrons completely fill the first two Brillouin zones. Presumably there is a gap between the second and third zones whose energy is the same as the gap between the bands deduced from band theory. We say "presumably" because the zone theory can only predict that a gap may exist. If it does exist, zone theory cannot predict its width.

In the zone theory, the bands arise as a consequence of the diffraction of electrons by the lattice; in the band theory, they are deduced from the splitting of states that occurs upon condensation of two or more atoms into a molecular system. In either case, the bands are identical. The zone theory permits a quantitative calculation of the distribution of states in energy,

enabling us to rationalize the Hume-Rothery rules of alloying and to account for the relation between the conductivity of metals and their valencies. The band theory permits us to make qualitative predictions more readily about the electronic structure of simple solids. Both theories have additional applications which will be developed later.

5.11 Ionic-covalent bonding and van der Waals bonding

When covalent bonds are formed between identical atoms, the electron orbitals will overlap in such a way that the electron distribution is symmetrical. However, when bonds are formed between different kinds of atoms, generally the electron distribution will be disturbed so that the centroid of the positive charge due to the nuclei will not coincide with the centroid of the electron density. In other terms, we could imagine that the bonding electrons spend a greater proportion of time in the neighborhood of one atom than in the neighborhood of the other. In this case the bond will have a permanent dipole moment, and the bond itself will have some ionic and some covalent character. Symmetry may still prevent the molecule as a whole from having a net dipole moment. For example, the C—H bonds in methane are all moderately polar, but they are centrosymmetric in space and the molecule will have no net dipole moment.

The polarity or lack of it in a molecule plays an important part in the intermolecular forces called *van der Waals forces.* These are relatively weak forces in comparison with the bonds between atoms; nevertheless, they are sufficiently strong to influence the boiling points of different compounds and they are the only forces that contribute to the bonding in the crystalline lattice of substances such as sugar, dry ice, etc. The stronger van der Waals forces always occur in polar molecules. However, even in the absence of a permanent dipole in the molecule, there will be very weak van der Waals forces arising from the transient dipoles associated with the motion of the electrons in an atom or molecule. The very low boiling points of substances such as hydrogen and helium are due to the weakness of the van der Waals forces that are operative.

Those molecules which do possess a permanent dipole moment will have a higher dielectric constant at frequencies that are low in comparison with the normal rotational frequency of the molecule. A few of these, such as chlorinated diphenyl and chlorinated naphthalene, are important as capacitor impregnants for this reason.

5.12 Ionic bonding

The extreme case where the bonding electrons spend all of the time associated with atoms of one kind and none of the time associated with atoms of

the other kind leads to the *ionic bond.* Practically all of the bond energy is due to the electrostatic attraction between the ions of opposite charge. This type of bond always leads to a crystalline structure in which ions of both kinds are incorporated into the lattice. There are no individual molecules in which the bond is entirely ionic. There are, however, ionic crystals in which the bonding is partially covalent.

The potential energy of ionic crystals is often expressed in terms of a quantity called the *Madelung constant.* This is a function of structure and ionic charge only and is based upon the assumption that the entire potential energy is electrostatic in origin. Since it is primarily used for purposes of comparison of the bonding energies of different crystals rather than for the computation of absolute bonding energies, the units are chosen for convenience in computing the value of the constant and do not represent the potential energy in common units. The potential energy of an ion pair is

$$V = \frac{e^+e^-}{r^2}$$

where the e's represent the charges on the ions and r is the distance between them. In computing the Madelung constant, the charge of an electron is taken as unity and r is measured in units based upon the distance between neighboring anions and cations. The negative charge is considered to be negative in a numerical sense. The total potential energy of a crystal is determined by summing over all combinations of ion pairs in the crystal:

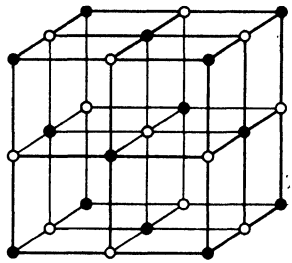
$$V_{tot} = -\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{e^+e^-}{r_{ij}}$$

This quantity depends upon the number of ions in the crystal, but if divided by the number of ion pairs involved in the summation, it approaches a limit fairly rapidly and then is independent of crystal size, and depends only upon the charge of the ions and the structure of the crystal. When the positive and negative ions in the crystal have the same charge, the definition given suffices, but when they differ, stoichiometric units must be substituted for ion pairs. The term *stoichiometric unit* refers to combinations such as CaF_2, Al_2O_3, $Na_3(PO_4)$, etc. Table 5.2 lists the Madelung constants of some common crystal types.

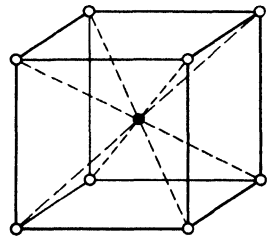
Table 5.2 *Madelung constants of common crystals*

Material	*Madelung constant*
Sodium chloride	1.7476
Cesium chloride	1.7627
Zincblende	1.6381
Fluorite	5.0388
Cuprite	4.1155
Rutile	4.816

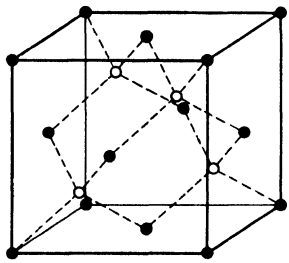
The materials listed in Table 5.2 are representative of a variety of crystal types. The Madelung constant, of course, does not depend upon the chemical



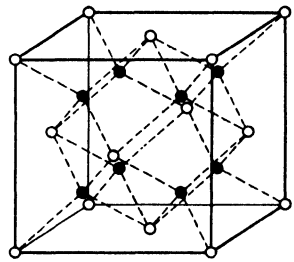
The NaCl structure



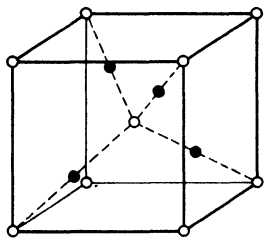
The cesium chloride, CsCl, structure



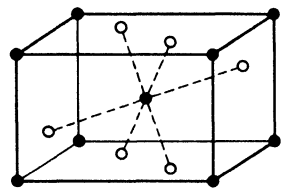
Zinc blende, ZnS, structure



The fluorite, CaF_2, structure



The cuprite, Cu_2O, structure



The rutile, TiO_2, structure

Figure 5.23 Lattices of the crystals listed in Table 5.2.

makeup of the crystal. Figure 5.23 shows the structure of each material given in the table. After Sherman, *Chem. Rev.*, **11**, 93 (1932).

PROBLEMS

5.1 Using the Pauli exclusion principle, construct the electronic structure of Chlorine (atomic number 17).

5.2 Compute the frequency of the radiation emitted when a hydrogen atom undergoes a transition from a $_2s$ state to a $_1s$ state.

5.3 Discuss the physical significance that is usually attached to the n, l, m, and s quantum numbers.

5.4 Discuss the formation of the valence and conduction bands in a covalent solid.

5.5 Discuss the nature of the covalent, polar, and ionic bonds.

5.6 What is meant by the term "hybridization" of bonds?

5.7 Compute the distance from the origin to the Brillouin plane in n-space that would represent lattice diffraction for a wave traveling in the $\langle 111 \rangle$ crystallographic direction in a bcc lattice.

5.8 Verify the Madelung constant for NaCl given in Table 5.3.

5.9 Discuss the concept of resonance or exchange energy and comment upon the pertinence of this concept in accounting for the metallic bond.

6

SYSTEMATIC CORRELATION OF PROPERTIES

In essence, a material property is a statement of the relationship between a change in the environment of a material and the response that this change elicits from the material itself. Within this rather broad definition there are several distinct categories which must be treated in distinctly different ways. Certain material properties are based upon qualitative concepts and can be defined only in terms of a specific test which purports to attach a number to the quality being measured. These numbers will rank different materials with respect to hardness, brittleness, ductility, and the like, but the numbers themselves will have a relative rather than an absolute meaning, and different tests will provide different scales of numbers to describe the same property. There is one fundamental factor which serves to distinguish properties of this type from any other category that we may define: in every case the test itself will leave the material in a permanently altered condition.

There are other categories of material properties which can be defined in absolute terms that are independent of the type of test devised to measure them. These categories are distinguished from one another upon the basis of both reversibility in the thermodynamic sense of the word and whether the system under discussion is in a steady state, a metastable state, or in a state of equilibrium. Before attempting to define any of these categories, it is necessary to discuss certain system parameters which are of importance in defining the categories as well as in defining the properties.

6.1 System parameters

When studying any system, its state may be completely specified by either of two sets of parameters. These are such that a pair, one from each set, when multiplied together will result in a product having the dimensions of energy per unit volume. Those parameters which may be measured at the surface of the material under consideration, or which are more closely associated with the environment of the material than with the material itself, are called *intensive parameters.* The other set of parameters is more closely associated with the material itself than with the environment; these are called *extensive parameters.*

The following list provides typical examples of both, together with the symbols used to represent them. The list is arranged so that parameters occupying the same position under each heading will be the pair whose product will have the dimensions of energy per unit volume.

Intensive parameters		*Extensive parameters*	
Temperature	T	Entropy per unit volume	S
Stress	σ	Strain	ϵ
Magnetic field intensity	H	Magnetic induction (per unit volume)	B
Electric field intensity	E	Electric displacement (per unit volume)	D
Pressure	P	Volume (relative to some standard volume)	ρ

6.2 Class I: equilibrium properties

If a system is in a state of equilibrium with its environment, then an incremental change in an intensive parameter will produce corresponding changes in all of the extensive parameters. If we select one of these and follow the changes it experiences as an intensive parameter is increased through several successive increments, allowing equilibrium to be established after each change, a curve such as that shown in Fig. 6.1 will result. If the intensive parameter is then decreased through successive increments to its original value, and if the extensive parameter varies in such a manner that the same curve is traced in the reverse direction, the system is said to be reversible with respect to these two parameters. If we select all of the parameters for which this reversibility holds, then each intensive parameter may be expressed as a function of all of the extensive parameters. Conversely, each of the extensive parameters may be expressed as a function of all of the intensive parameters.

For example, a change in entropy would be related to changes in tem-

perature, pressure, etc. by the equation

$$dS = \frac{\partial S}{\partial T}\Big)_{P,E} dT + \frac{\partial S}{\partial P}\Big)_{T,E} dP + \frac{\partial S}{\partial E}\Big)_{T,P} dE + \ldots \tag{6.1}$$

Arbitrary variations in both sets of parameters are related to one another by a

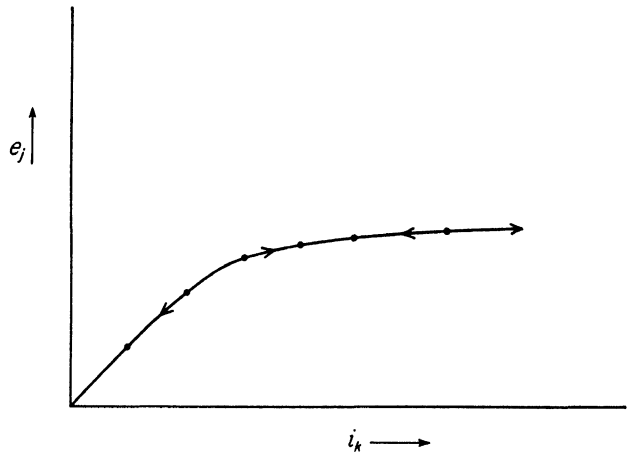


Figure 6.1 When a single functional relationship exists between some intensive parameter i_k and some extensive parameter e_j, the system will be thermodynamically reversible with respect to these parameters. Under these circumstances, the response of the system to an environmental change is characterized by the rate of change of these parameters with respect to one another. These partial derivatives are the equilibrium properties of the material.

set of equations similar to Eq. (6.1). These are concisely represented by a single matrix equation

$$(de_j) = (C_{j,k})(di_k) \tag{6.2}$$

The elements of the column matrix (de_j) are small changes in the extensive parameters, and the column matrix (di_k) represents the changes of the intensive parameters. The elements $C_{j,k}$ follow the notation used in most mathematics texts, i.e.

$$\frac{\partial e_j}{\partial i_k}\Big)_{e_l,e_m\ldots=\text{constant}}$$

Equation (6.2) provides the formal definition of equilibrium material properties. Following Zwikker,[1] we call the quantities $C_{j,k}$ *compliances*. It is very

[1] C. Zwikker, *Physical Properties of Solid Materials* (London: Pergamon Press, 1954), chap. 5.

important to recognize that the definition of a partial derivative requires that $C_{j,k}$ be the rate of change of the jth extensive parameter with respect to the kth intensive parameter that occurs when *all of the other intensive parameters are held constant.*

To make these ideas more concrete, let us consider as general a situation as is likely to be encountered. This would involve a system in which all of the parameters listed on p. 116 would be important; the resulting compliance matrix would be

$$\begin{bmatrix} C_{S,T} & C_{S,H} & C_{S,E} & C_{S,P} & C_{S,\sigma} \\ C_{B,T} & C_{B,H} & C_{B,E} & C_{B,P} & C_{B,\sigma} \\ C_{D,T} & C_{D,H} & C_{D,E} & C_{D,P} & C_{D,\sigma} \\ C_{\rho,T} & C_{\rho,H} & C_{\rho,E} & C_{\rho,P} & C_{\rho,\sigma} \\ C_{\epsilon,T} & C_{\epsilon,H} & C_{\epsilon,E} & C_{\epsilon,P} & C_{\epsilon,\sigma} \end{bmatrix}$$

Some of these compliances will be familiar to every reader; others will be similar to but not identical with properties that every student of engineering or physics is familiar with; and still others belong to important but highly specialized branches of physics or engineering and will not be familiar to the general reader. Almost all of them will be treated in some detail later, and for the present it will suffice to comment upon a select few to bring them into focus.

The compliance $C_{S,T}$ is the rate of change of entropy per unit volume at constant pressure, electric field intensity, etc. This is closely related to the familiar specific heat at constant pressure. The exact relationship is

$$C_p = \frac{\rho}{T} C_{S,T}$$

The elastic compliance $C_{\epsilon,\sigma}$ looks very much like the inverse of the modulus of elasticity. There is a technical difference, which is discussed in Chap. 8. The compliance $C_{D,E}$ is the dielectric constant of the material, provided that it is understood the measurement is made isothermally.

Among the cross terms of the compliance matrix are $C_{\epsilon,T}$, which is the familiar coefficient of thermal expansion. Less familiar is the compliance $C_{D,\sigma}$ which implies that it is possible to produce an electric displacement by stressing a material. This does occur in certain materials, and as a matter of fact the phenomenon characterized by this compliance is quite important technically. This phenomenon is called *piezoelectricity* and $C_{D,\sigma}$ is called the *piezoelectric stress coefficient.*

We could just as well have defined a set of equilibrium properties in the form

$$R_{k,j} = \frac{\partial i_k}{\partial e_j}\bigg)_{i_1,i_2,\ldots = \text{constant}}$$

Quantities such as these are called *rigidities*. Again, small changes in extensive and intensive parameters could be related through a matrix equation

$$(di_k) = (R_{k,j})(de_j) \tag{6.3}$$

Comparison between Eqs. (6.2) and (6.3) makes evident the relationship

$$(R_{k,j}) = (C_{j,k})^{-1} \tag{6.4}$$

The rigidity matrix in extended form would be

$$\begin{bmatrix} R_{T,S} & R_{T,B} & R_{T,D} & R_{T,\rho} & R_{T,\epsilon} \\ R_{H,S} & R_{H,B} & R_{H,D} & R_{H,\rho} & R_{H,\epsilon} \\ R_{E,S} & R_{E,B} & R_{E,D} & R_{E,\rho} & R_{E,\epsilon} \\ R_{P,S} & R_{P,B} & R_{P,D} & R_{P,\rho} & R_{P,\epsilon} \\ R_{\sigma,S} & R_{\sigma,B} & R_{\sigma,D} & R_{\sigma,\rho} & R_{\sigma,\epsilon} \end{bmatrix}$$

where $$R_{T,S} = \frac{\partial T}{\partial S}\Big)_{B,D,\rho,\epsilon = \text{constant}}$$

and $$R_{\sigma,B} = \frac{\partial \sigma}{\partial B}\Big)_{S,D,\rho,\epsilon = \text{constant}}$$

Among the diagonal elements are terms such as $R_{\sigma,\epsilon}$ the isentropic elastic modulus, $R_{P,\rho}$ the inverse of the isentropic bulk modulus. Among the nondiagonal terms are quantities such as the pyroelectric constant $R_{T,D}$.

Both the rigidity matrix and the compliance matrix are symmetric, and

$$C_{i,j} = C_{j,i}, \qquad R_{i,j} = R_{j,i}$$

This fact is of both theoretical and practical importance, since one of the pairs of equal quantities may be determined much more readily than the other. This assertion may be proven by the following thermodynamic arguments. Let du be a small change in internal energy due to a change in the jth extensive parameter:

$$du = i_j\, de_j$$

Thus $$i_j = \frac{\partial u}{\partial e_j}\Big)_{e_k e_l, \ldots = \text{constant}}$$

Differentiating gives

$$\frac{\partial i_j}{\partial e_k} = \frac{\partial^2 u}{\partial e_j\, \partial e_k} \tag{6.5}$$

Since u is a state function, the order of differentiation is immaterial, hence

$$\frac{\partial i_j}{\partial e_k} = \frac{\partial^2 u}{\partial e_j\, \partial e_k} = \frac{\partial^2 u}{\partial e_k\, \partial e_j} = \frac{\partial i_k}{\partial e_j} \tag{6.6}$$

Consequently $$R_{j,k} = R_{k,j} \tag{6.7}$$

The proof that $C_{j,k} = C_{k,j}$ follows similar lines except that the free energy function F is used instead of the internal energy u.

Since the compliance matrix is symmetric, we may write

$$C_{S,H} = C_{B,T}, \qquad C_{S,\sigma} = C_{\epsilon,T}, \qquad C_{\epsilon,H} = C_{B,\sigma}$$

Thus the piezomagnetic stress coefficient $C_{B,T}$ is equal to the magnetostrictive coefficient $C_{\epsilon,H}$. The coefficient of thermal expansion $C_{\epsilon,T}$ is identical with the entropy change accompanying an isothermal application of stress $C_{S,\sigma}$. Finally, if a change in temperature is capable of producing a magnetic induction (this phenomenon is called *pyromagnetism*) the magnitude of this effect is equal to the change in entropy produced by a magnetic field.

In the foregoing discussion we have used the intensive parameters E, T, H, and σ as examples. This does not exhaust the possible intensive parameters that may be of importance in determining the state of a system, but in most cases fewer parameters will suffice. In general, the question of which parameters must be considered in a problem depends upon what properties are to be evaluated and what conditions apply in the evaluation.

For example, if we wish to measure only the elastic properties of a material we may exclude electric and magnetic fields from the sample at the time of measurement and may neglect atmospheric pressure since it would remain constant. This reduces the number of intensive parameters to be specified to two: stress and temperature. Therefore, in this case, the entire compliance matrix is

$$\begin{bmatrix} C_{S,T} & C_{S,\sigma} \\ C_{\epsilon,T} & C_{\epsilon,\sigma} \end{bmatrix}$$

Because the matrix is symmetric, $C_{\epsilon,T} = C_{S,\sigma}$, and since $C_{\epsilon,T}$ is simply the thermal expansion tensor, we rewrite this matrix in the form

$$\begin{bmatrix} C_{S,T} & \alpha \\ \alpha & C_{\epsilon,\sigma} \end{bmatrix}$$

This bears an inverse relationship to the rigidity matrix

$$\begin{bmatrix} R_{T,S} & R_{T,\epsilon} \\ R_{\sigma,S} & R_{\sigma,\epsilon} \end{bmatrix}$$

where again $R_{T,\epsilon} = R_{\sigma,S} = \lambda$; λ is the isentropic temperature rise accompanying a strain.

Now we wish to relate λ to α and $C_{\epsilon,\sigma}$ to $R_{\sigma,\epsilon}$. From the definition of the inverse of a matrix we have

$$C_{ij} = \frac{\text{Cofactor } R_{ij}}{\text{Det } R} \tag{6.7a}$$

and

$$R_{ij} = \frac{\text{Cofactor } C_{ij}}{\text{Det } C} \tag{6.7b}$$

In the following it is to be understood that the operations are purely symbolic, since the elements concerned may be tensors of any rank.

$$\begin{aligned} C_{\epsilon,\sigma} &= \frac{R_{T,S}}{R_{\sigma,\epsilon}R_{T,S} - \lambda^2} && \text{(a)} \\ R_{\sigma,\epsilon} &= \frac{C_{S,T}}{C_{S,T}C_{\epsilon,\sigma} - \alpha^2} && \text{(b)} \\ \alpha &= \frac{-\lambda}{R_{T,S}R_{\sigma,\epsilon} - \lambda^2} && \text{(c)} \\ \lambda &= \frac{-\alpha}{C_{S,T}C_{\epsilon,\sigma} - \alpha^2} && \text{(d)} \end{aligned} \tag{6.8}$$

and

It should be emphasized that a rigidity is *not* the reciprocal of a compliance. The only way compliances and rigidities may be related to one another is by forming the inverse of either matrix according to the methods just used. This point is often the source of confusion, since many material properties are commonly expressed in terms of the reciprocal of a rigidity and it seems to be a normal tendency to identify these (improperly) with the corresponding compliance. For example, $-1/\partial\rho/\partial P\Big)_S$ is an adiabatic bulk modulus and is the negative reciprocal of a rigidity, whereas $\partial\rho/\partial P\Big)_T$ is a true compliance and is determined under conditions where the system is kept at constant temperature rather than at constant entropy. Again the adiabatic elastic modulus of a material $\partial\sigma/\partial\epsilon\Big)_S$ would normally be evaluated from the speed of sound in a material, since under these circumstances the strain at a point occurs so rapidly that the process is essentially isentropic. On the other hand, normal usage of the term *elastic modulus* refers to the reciprocal of the compliance $\partial\sigma/\partial\epsilon\Big)_{T,H,E}$ which is measured at constant temperature in a static test.

6.3 Mathematical nature of compliances and rigidities

At this point it will be useful to consider the mathematical nature of some of the equilibrium properties that constitute the elements of the matrix (C_{ij}). Again considering a system whose state is specified by the parameters S, D,

B, and ϵ, the properties associated with these parameters are defined by the set of symbolic equations:

$$\begin{aligned} dS &= C_{S,T}\,dT + C_{S,E}\,dE + C_{S,H}\,dH + C_{S,\sigma}\,d\sigma \quad \text{(a)} \\ dD &= C_{D,T}\,dT + C_{D,E}\,dE + C_{D,H}\,dH + C_{D,\sigma}\,d\sigma \quad \text{(b)} \\ dB &= C_{B,T}\,dT + C_{B,E}\,dE + C_{B,H}\,dH + C_{B,\sigma}\,d\sigma \quad \text{(c)} \\ d\epsilon &= C_{\epsilon,T}\,dT + C_{\epsilon,E}\,dE + C_{\epsilon,H}\,dH + C_{\epsilon,\sigma}\,d\sigma \quad \text{(d)} \end{aligned} \tag{6.9}$$

Among the extensive parameters, we find a scalar quantity dS, two vectors $d\mathbf{D}$ and $d\mathbf{B}$, and a quantity $d\boldsymbol{\varepsilon}$ which is a tensor of rank two. The intensive parameters include a scalar dT, two vectors $d\mathbf{E}$ and $d\mathbf{H}$, and a tensor of rank two $d\boldsymbol{\sigma}$.

We may deduce the nature of the coefficients by recognizing that all terms in each equation must be of the same nature. In the first, all must be scalars. Therefore, $C_{S,T}$ must be a scalar since dT is a scalar. The term which was previously written as $C_{S,E}\,dE$ must be the dot product of two vectors, $\mathbf{C}_{S,E}$ and $d\mathbf{E}$, since the electric field intensity is itself a vector. Similarly $\mathbf{C}_{S,H}$ must be a vector. In the last term of Eq. (6.9a) the stress $d\boldsymbol{\sigma}$ is a tensor of rank two. Since the term itself is a scalar, $C_{S,\sigma}\,d\sigma$ must represent a double contraction between two rank two tensors which would be written $\mathbf{C}_{S,\sigma} : d\boldsymbol{\sigma}$.

In Eq. (6.9b), dD, the change in dielectric displacement per unit of volume, is a vector. Consequently the terms $C_{D,T}\,dT$, $C_{D,E}\,dE$, $C_{D,H}\,dH$, and $C_{D,\sigma}\,d\sigma$ must also be vectors. We then deduce that $\mathbf{C}_{D,T}$ is a vector because dT is a scalar. The dot product of the tensor of rank two, $\mathbf{C}_{D,E}$, and the vector $d\mathbf{E}$ produces a vector $\mathbf{C}_{D,E}\,dE$. $\mathbf{C}_{D,\sigma}$ is a tensor of rank three and the term $C_{D,\sigma}\,d\sigma$ represents a contraction between the two tensors. In a similar manner, we may deduce that $\mathbf{C}_{E,H}$ and $\mathbf{C}_{\epsilon,\sigma}$ are rank two and rank three tensors respectively. The nature of all the remaining terms can be deduced from the symmetry of the compliance matrix.

Let us return to Eqs. (6.8). The following comments apply to the entire set, but let us select (6.8d) for discussion. The quantities α and λ are rank two tensors and $C_{\epsilon,\sigma}$ is a rank four tensor. We have not defined an operation equivalent to division for tensors. Consequently, Eq. (6.8d) must be interpreted in accordance with defined tensor operations. We first write the equation in the form

$$\lambda(C_{S,T}C_{\epsilon,\sigma} - \alpha^2) = -\alpha$$

then take cognizance of the rank of any tensors present, and introduce appropriate contractions or dyadic products to make the equation homogeneous in the rank of its tensors. Thus it may be rewritten in the form

$$\boldsymbol{\lambda} : (C_{S,T}\mathbf{C}_{\epsilon,\sigma} - \boldsymbol{\alpha}^2) = -\boldsymbol{\alpha}$$

where $\boldsymbol{\alpha}^2$ is the dyadic product $\boldsymbol{\alpha}\boldsymbol{\alpha}$. The quantities on each side of the equation are now tensors of rank two.

To clarify the ideas presented thus far we will consider certain natural phenomena that the compliances or rigidities describe.

Vector properties: pyroelectricity, pyromagnetism. Certain crystals when heated will assume a surface charge. Actually crystals such as these are polarized at any temperature, but the surface charge is quickly neutralized by atmospheric leakage and only becomes apparent during a transient period following a temperature change. This phenomenon is called *pyroelectricity*. It occurs only in ionic crystals which lack a center of symmetry. Such crystals will have an asymmetric distribution of positive and negative ions, which produces a net polarization in some one direction. A strain caused by a temperature variation would change the magnitude of the polarization and cause the polarity of the crystal to become manifest.

The *electrocaloric effect* is the inverse phenomenon in which a pyroelectric crystal will evolve heat when it is placed in an electric field. Since $\partial S/\partial \mathbf{E}\big)_{T,H}$ equals $\partial \mathbf{D}/\partial T\big)_{E,H}$, either compliance provides a quantitative description of this phenomenon. If these are constant, the heat liberated at constant temperature could be computed from

$$q = T\mathbf{C}_{S,E} \cdot \Delta\mathbf{E} = T\mathbf{C}_{D,T} \cdot \Delta\mathbf{E}$$

The change in polarization due to a temperature change would be

$$\Delta\mathbf{D} = \mathbf{C}_{S,E}\,\Delta T = \mathbf{C}_{D,T}\,\Delta T$$

Each of the compliances is associated with a particular direction called the *pyroelectric axis*. If $\mathbf{E}$ is perpendicular to this axis, no heat will be liberated. Also, changes in T produce a polarization in a direction parallel to this axis. In Chap. 7 we will establish the relationship of this axis to the crystallographic axes.

The internal electric field produced by a temperature rise and the isentropic temperature change due to a change in polarization are computed from either of the rigidities $\partial \mathbf{E}/\partial S\big)_{D,B,\sigma}$ or $\partial T/\partial \mathbf{D}\big)_{S,B,\sigma}$. Thus

$$dT = \frac{q}{T}\mathbf{R}_{T,D} \cdot d\mathbf{D} = \frac{q}{T}\mathbf{R}_{E,S} \cdot d\mathbf{D}$$

whereas

$$\Delta\mathbf{E} = \mathbf{R}_{E,S} \int \frac{q}{T}\,dT = \mathbf{R}_{T,D} \int \frac{q}{T}\,dT$$

A similar phenomenon occurs in connection with a magnetic field. Certain crystals will develop a magnetic moment as a consequence of a tem-

perature change. Conversely if they are placed in a magnetic field, heat will be liberated. The coefficient $\mathbf{C}_{S,H}$ is called the *magnetocaloric compliance* and is, of course, equal to the pyromagnetic compliance $\mathbf{C}_{B,T}$.

The pyromagnetic effect occurs because some of the ions in the crystal have one or more unpaired electrons and have a net magnetic moment. In a magnetic field, they would tend to align with the field and would consequently lose some of their energy of motion, thus causing a liberation of heat.

Scalar-tensor effects. Thermal expansion is a typical scalar-tensor effect, the strain components being related to the temperature change and the thermal expansion tensor by

$$\Delta\boldsymbol{\varepsilon} = \mathbf{C}_{\epsilon,T}\,\Delta T$$

The thermal expansion tensor is a symmetric tensor of rank two. As with all symmetric tensors of rank two, a set of principal axes can be found that may be characterized by, at the most, three independent coefficients. In a later chapter we will show how the principal axes of this tensor are oriented with respect to the crystallographic axes. We shall also show that any tensor property of rank two becomes isotropic for cubic symmetry and has two independent coefficients when the crystal symmetry is higher than orthorhombic but lower than cubic.

The complementary coefficient $\mathbf{C}_{S,\sigma}$ describes the heat liberated by an isothermal stress. This also is a tensor of rank two, and the operation which related the heat liberated to the applied stress is a contraction between $\mathbf{C}_{S,\sigma}$ and $d\boldsymbol{\sigma}$. If we hold T, $\mathbf{D}$, $\mathbf{H}$, etc. constant, then the heat liberated by an isothermal stress is

$$\frac{q}{T} = \Delta S = \mathbf{C}_{S,\sigma} : \boldsymbol{\sigma} = \alpha_{ij}\sigma_{ij}$$

Vector-tensor effects. The coefficients $\mathbf{C}_{D,E}$, $\mathbf{C}_{D,H}$, $\mathbf{C}_{B,H}$ are also tensor properties of rank two, but they operate in a somewhat different manner than the thermal expansion tensor. A change in a vector intensive parameter will be related to a change in a vector extensive parameter through a contraction of the vector intensive parameter and the rank two tensor property. In the case of dielectric permittivity, an electric field $\mathbf{E}$ will produce an electric displacement $\mathbf{D}$ according to the equation

$$d\mathbf{D} = \mathbf{C}_{D,E} \cdot d\mathbf{E}$$

Again the principal axes of the permittivity tensor will bear some definite relationship to the crystallographic axes which will primarily be determined by crystal symmetry. The tensor itself would be characterized by three independent coefficients. If we chose a coordinate system oriented along the

principal axes of the permittivity tensor, we could write

$$dD_i = \frac{\partial D_i}{\partial E_k} dE_k$$

Ordinarily, then, the direction of the displacement vector would not coincide with that of the electric field.

Tensor-tensor effects. Those crystals which are pyroelectric are also piezoelectric. That is, when stressed, they become polarized; conversely, they will suffer a deformation when placed in an electric field. This converse effect is called *electrostriction.* Assuming that the compliances are constant, then the application of a stress, with all other intensive parameters held constant, will produce a dielectric displacement

$$\mathbf{D} = \mathbf{C}_{D,\sigma} : \boldsymbol{\sigma}$$

The compliance $\mathbf{C}_{D,\sigma}$ is a symmetric tensor of rank three, and a double contraction with the stress tensor produces the displacement vector $\mathbf{D}$. The piezoelectric compliance has a total of 27 components but because it is symmetric only 18 of these are distinct. If an arbitrary coordinate system is selected, the dielectric displacements along axes x_1, x_2, and x_3 could be related to the tensile and shearing stresses in this coordinate system by the set

$$\begin{aligned}
D_1 &= d_{11}\sigma_{11} + d_{12}\sigma_{22} + d_{13}\sigma_{33} + d_{14}\sigma_{23} + d_{15}\sigma_{13} + d_{16}\sigma_{12} \\
D_2 &= d_{21}\sigma_{11} + d_{22}\sigma_{22} + d_{23}\sigma_{33} + d_{24}\sigma_{23} + d_{25}\sigma_{13} + d_{26}\sigma_{12} \\
D_3 &= d_{31}\sigma_{11} + d_{32}\sigma_{33} + d_{33}\sigma_{33} + d_{34}\sigma_{23} + d_{35}\sigma_{13} + d_{36}\sigma_{12}
\end{aligned}$$

The coefficients d_{ij} can be related to the components of the tensor $\mathbf{C}_{D,\sigma}$ by expanding the equation

$$\mathbf{D} = \mathbf{C}_{D,\sigma} : \boldsymbol{\sigma}$$

The components of the tensor $\mathbf{C}_{D,\sigma}$ may be represented by C_{ijk} and those of the tensor $\boldsymbol{\sigma}$ by σ_{jk}. Then

$$D_i = C_{ijk}\sigma_{jk}$$

For $i = 1$, we may write

$$D_1 = C_{111}\sigma_{11} + C_{121}\sigma_{21} + C_{131}\sigma_{31} + C_{122}\sigma_{22} + C_{132}\sigma_{32} + C_{113}\sigma_{13} + C_{112}\sigma_{12} + C_{123}\sigma_{23} + C_{133}\sigma_{33}$$

But because both $\boldsymbol{\sigma}$ and $\mathbf{C}_{D,\sigma}$ are symmetric, $C_{121} = C_{112}$ and $\sigma_{21} = \sigma_{12}$, etc. Hence

$$\begin{aligned}
d_{11} &= C_{111}, & d_{14} &= 2C_{123} = 2C_{132} \\
d_{12} &= C_{122}, & d_{15} &= 2C_{113} = 2C_{131} \\
d_{13} &= C_{133}, & d_{16} &= 2C_{112} = 2C_{121}
\end{aligned}$$

In a similar manner, the relationships between the other tensor components and the coefficients d_{ij} could be deduced.

Second-order effects. As a rule, the compliances and rigidities do not depend strongly upon the magnitudes of the intensive and extensive parameters, and may be considered constant over restricted ranges of these parameters. However, there are cases in which this is not true. In these instances the second-order effects may be so great that they mask the first-order effect. As a case in point, we may consider the behavior of the magnetocaloric effect in ferromagnetic materials near the Curie temperature. Using T and $\mathbf{H}$ as the intensive parameters which specify the system, we may write

$$dS = \frac{\partial S}{\partial T}\Big)_B dT + \frac{\partial S}{\partial \mathbf{H}}\Big)_T \cdot d\mathbf{H} = C_{S,T}\, dT + \mathbf{C}_{S,H} \cdot d\mathbf{H} \qquad (6.10)$$

If $\mathbf{H}$ is changed isothermally

$$\int dS = \int \mathbf{C}_{S,H} \cdot d\mathbf{H}$$

Using a Taylor's expansion to evaluate the integral, we have

$$\Delta S = C_{S,H} \cdot \Delta\mathbf{H} + \frac{1}{2}\frac{\partial \mathbf{C}_{S,H}}{\partial \mathbf{H}} : \Delta\mathbf{H}^2 + \ldots \qquad (6.11)$$

Note that $\mathbf{H}^2$ is treated as a dyad and $\partial \mathbf{C}_{S,H}/\partial \mathbf{H}$ as a tensor of rank two.[2]

Since $\mathbf{C}_{S,H} = \dfrac{\partial S}{\partial \mathbf{H}}\Big)_T = \dfrac{\partial \mathbf{B}}{\partial T}\Big)_{\mathbf{H}} = \mathbf{C}_{B,T}$,

$$\frac{\partial \mathbf{C}_{S,H}}{\partial \mathbf{H}} = \frac{\partial^2 \mathbf{B}}{\partial \mathbf{H}\, \partial T} = \frac{\partial \mathbf{C}_{B,H}}{\partial T}$$

Consequently,

$$\Delta S = \mathbf{C}_{S,H} \cdot \Delta\mathbf{H} + \frac{1}{2}\frac{\partial \mathbf{C}_{B,H}}{\partial T} : \Delta\mathbf{H}^2 + \ldots$$

This may be rewritten in the form

$$S - S_0 = \mathbf{C}_{S,H} \cdot (\mathbf{H} - \mathbf{H}_0) + \frac{1}{2}\frac{\partial C_{B,H}}{\partial T} \cdot (\mathbf{H} - \mathbf{H}_0)^2 + \ldots \qquad (6.12)$$

If we choose S_0 as being the entropy per unit volume when $\mathbf{H}_0 = 0$, and if $\mathbf{C}_{S,H}$ and $\partial \mathbf{C}_{B,H}/\partial T$ are constant in the range $\mathbf{H} = 0$ to $\mathbf{H}$, differentiation produces

$$dS = \mathbf{C}_{S,H} \cdot d\mathbf{H} + \frac{\partial \mathbf{C}_{B,H}}{\partial T} : \mathbf{H}\, d\mathbf{H} \qquad (6.13)$$

[2] See Chap. 3, Sec. 6, for an explanation of why this is a tensor of rank two.

The term $(\partial \mathbf{C}_{B,H}/\partial T) : \mathbf{H}\, d\mathbf{H}$ describes the quadratic magnetocaloric effect. In the neighborhood of the Curie temperature, $\partial \mathbf{C}_{B,H}/\partial T$ may be quite large, in which case an increase in magnetic field strength will produce an evolution of heat, which for small changes in $\mathbf{H}$ will be proportional to the square of the magnetic field intensity.

In a similar manner, we could have deduced that a change in temperature would produce a quadratic change in magnetic induction:

$$d\mathbf{B} = \mathbf{C}_{B,T}\, dT + \frac{\partial C_{S,T}}{\partial \mathbf{H}}\, T\, dT + \ldots$$

The temperature change would result in a linear change in $\mathbf{B}$ proportional to the pyromagnetic compliance and a quadratic change, which would be proportional to the rate at which the thermal capacitance changes with magnetic field intensity.

The previous development was not the most general, and we may inquire how the entropy per unit volume and the magnetic induction change when both $\mathbf{H}$ and T are varied. In this case

$$dS = C_{S,T}\, dT + \mathbf{C}_{S,H} \cdot d\mathbf{H}$$
$$d\mathbf{B} = \mathbf{C}_{B,T}\, dT + \mathbf{C}_{B,H} \cdot d\mathbf{H}$$

Without presenting the details, a Taylor's expansion followed by a differentiation gives

$$dS = C_{S,T}\, dT + \mathbf{C}_{S,H} \cdot d\mathbf{H} + \frac{\partial C_{S,T}}{\partial T}\, T\, dT + \frac{\partial \mathbf{C}_{B,H}}{\partial T} : \mathbf{H}\, d\mathbf{H} + \ldots \quad \text{(a)} \tag{6.14}$$

$$d\mathbf{B} = \mathbf{C}_{B,T}\, dT + \mathbf{C}_{B,H} \cdot d\mathbf{H} + \frac{\partial C_{S,T}}{\partial \mathbf{H}}\, T\, dT + \frac{\partial \mathbf{C}_{B,H}}{\partial \mathbf{H}} \cdot \mathbf{H}\, d\mathbf{H} + \ldots \quad \text{(b)}$$

This may be written in the matrix form

$$\begin{bmatrix} dS \\ d\mathbf{B} \end{bmatrix} = \begin{bmatrix} C_{S,T} + T\dfrac{\partial C_{S,T}}{\partial T} & \mathbf{C}_{S,H} + \mathbf{H} \cdot \dfrac{\partial \mathbf{C}_{B,H}}{\partial T} \\ \mathbf{C}_{B,T} + T\dfrac{\partial \mathbf{C}_{S,T}}{\partial \mathbf{H}} & \mathbf{C}_{B,H} + \mathbf{H} \cdot \dfrac{\partial \mathbf{C}_{B,H}}{\partial \mathbf{H}} \end{bmatrix} \begin{bmatrix} dT \\ d\mathbf{H} \end{bmatrix} \tag{6.15}$$

Again, the first matrix on the right side of the equation must be symmetric and we may write

$$\mathbf{C}_{s,H} + \mathbf{H} \cdot \frac{\partial \mathbf{C}_{B,H}}{\partial T} = \mathbf{C}_{B,T} + T\frac{\partial C_{S,T}}{\partial \mathbf{B}} \tag{6.16}$$

But $\mathbf{C}_{S,H} = \mathbf{C}_{B,T}$, therefore

$$\mathbf{H} \cdot \frac{\partial \mathbf{C}_{B,H}}{\partial T} = T\frac{\partial C_{S,T}}{\partial \mathbf{B}} \tag{6.17}$$

These ideas may be generalized for all second-order effects occurring in systems that require only two parameters to specify their equilibrium state. Let α and β represent any two intensive parameters and a and b represent their corresponding extensive parameters. The modulus $\partial a/\partial\alpha\Big)_\beta$ is represented as $C_{a,\alpha}$. Then the system properties, including the quadratic terms, are related by the matrix equation

$$\begin{bmatrix} da \\ db \end{bmatrix} = \begin{bmatrix} C_{a,\alpha} + \alpha \dfrac{\partial C_{a,\alpha}}{\partial \alpha} & C_{a,\beta} + \beta \dfrac{\partial C_{b,\beta}}{\partial \alpha} \\ C_{b,\alpha} + \alpha \dfrac{\partial C_{a,\alpha}}{\partial \beta} & C_{b,\beta} + \beta \dfrac{\partial C_{b,\beta}}{\partial \beta} \end{bmatrix} \begin{bmatrix} d\alpha \\ d\beta \end{bmatrix} \tag{6.18}$$

Using the symmetry of the matrix and the fact that $C_{a,\beta} = C_{b,\alpha}$ we find

$$\alpha \frac{\partial C_{a,\alpha}}{\partial \beta} = \beta \frac{\partial C_{b,\beta}}{\partial \alpha} \tag{6.19}$$

We shall have occasion later to consider these relationships in specific second-order effects in a treatment of ferromagnetism and ferroelectricity.

6.4 Class II: properties of steady-state systems

Thus far we have considered only those properties which are measurable by changing a system from one equilibrium state to another only infinitesimally different from the first. Now let us consider the properties which can be measured in a system not in equilibrium but in a steady state. Under these circumstances, it is meaningless to talk about its thermodynamic state. Instead, the extensive and intensive parameters will have a gradient at every point in the system. These gradients may induce a flow of some quantity such as heat, mass, or electric charge. If this should occur, the rate and direction of flow would depend not only upon the magnitudes and directions of the gradients causing it, but also upon the material itself. The quantities which relate the gradients to the flow rates are the steady-state properties of the material. Before defining these, we must characterize the flow rates themselves. To do this, we must state the direction of flow and the rate of flow through a unit area perpendicular to the direction of flow. All of this information can be embodied in one vector quantity, called *current density*, and usually symbolized by a $\mathbf{J}$. We distinguish mass current density, entropy current density, and electrical current density by subscripts m, s, and e respectively.

According to this definition, the rate of flow through a plane surface $\mathbf{A}$ would be $\mathbf{J} \cdot \mathbf{A}$ and the total quantity that would pass through the plane surface in time t would be $\mathbf{J} \cdot \mathbf{A}t$. The total rate of flow through a curved surface could be found by evaluating the surface integral $\int_\Sigma \mathbf{J} \cdot d\mathbf{A}$. If at every

point in the system there is an electric potential gradient, a temperature gradient, and a concentration gradient, then the current densities and the gradients would be related by a set of equations

$$-\mathbf{J}_m = \mathbf{C}_{m,c} \cdot \nabla c + \mathbf{C}_{m,V} \cdot \nabla V + \mathbf{C}_{m,T} \cdot \nabla T \quad \text{(a)}$$
$$-\mathbf{J}_e = \mathbf{C}_{e,c} \cdot \nabla c + \mathbf{C}_{e,V} \cdot \nabla V + \mathbf{C}_{e,T} \cdot \nabla T \quad \text{(b)} \quad (6.20)$$
$$-\mathbf{J}_s = \mathbf{C}_{s,c} \cdot \nabla c + \mathbf{C}_{s,V} \cdot \nabla V + \mathbf{C}_{s,T} \cdot \nabla T \quad \text{(c)}$$

In more general terms, we might write this set in the form of a matrix equation

$$(\mathbf{J}_\alpha) = -(C_{\alpha,\beta})(\nabla \varphi_\beta)$$

The matrix elements $C_{\alpha,\beta}$ are properties of the material under consideration and are called *conductivities*. The conductivities are tensors of rank two and the products between these and the gradients are dot products.

We also note when the nondiagonal terms are not zero, a flow of heat could be produced in the absence of a temperature gradient by one or both of the other gradients. Similarly a flow of electric current could be produced by a temperature gradient even though no electric field was present.

Just as we may compute the current densities from the gradients and the conductivities, we may also compute the gradient from the current densities. Thus

$$-(\nabla \varphi_\beta) = (\mathbf{C}_{\alpha,\beta})^{-1}(\mathbf{J}_\alpha) = (\varrho_{\beta,\alpha})(\mathbf{J}_\alpha) \quad (6.21)$$

The matrix elements $\rho_{\beta,\alpha}$ are called *resistivities*. These are also tensors of rank two. These relationships will be used extensively in a later chapter devoted to thermoelectric energy conversion and related topics.

6.5 Class III properties: systems with hysteresis

When a system exhibits extensive hysteresis in the relationship between a pair of parameters, properties defined in terms of these parameters cannot be incorporated into the formal relationships that we have derived for compliances, rigidities, conductivities, and resistivities. There are two reasons for this: the system is not thermodynamically reversible, and the parameters involved cannot be connected to one another by a unique functional relationship. This latter factor even makes it impossible to define a property as a compliance or resistivity without imposing some rather artificial verbal qualifications upon the definition.

An appreciation of the problems that arise in dealing with hysteresis can be acquired by considering in detail the behavior of a typical case. For our example we choose the relationship between **B** and **H** in a ferromagnetic material. Since both **B** and **H** are vectors, let us consider only the relationship between parallel components B_i and H_i and let us begin with zero

residual magnetism in the sample and zero magnetic field intensity. Our starting point, then, is the origin of the coordinate system of Fig. 6.2. If we increase H_i monotonically to a value aH_i, B_i and H_i will form the curve *oa*. If H_i is then decreased monotonically to a value $-aH_i$, the system will form the lefthand branch of the hysteresis loop to point *b*. Changing H_i

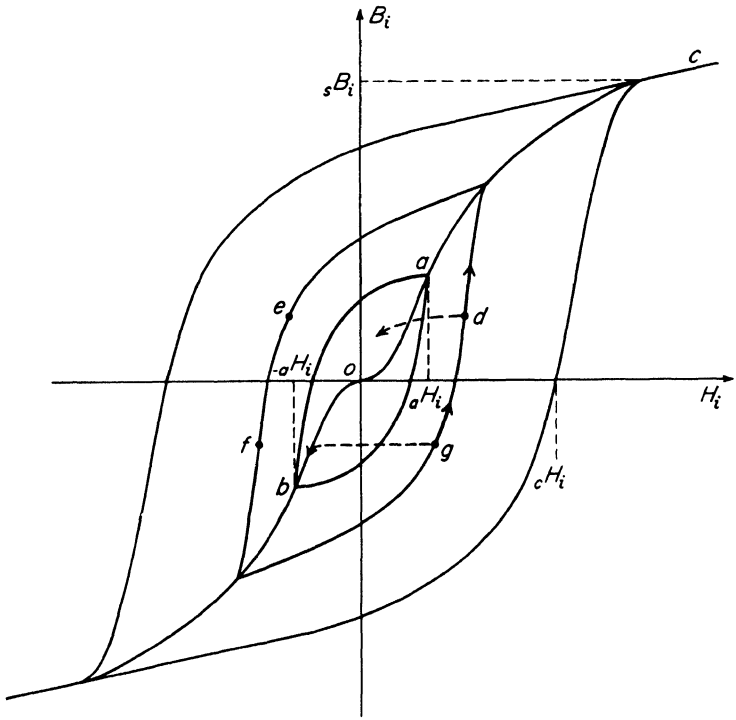


Figure 6.2 Hysteresis loop typical of Class III properties.

monotonically back to aH_i causes the righthand side of the hysteresis loop to be formed.

If instead of having initial values of zero, B_i and H_i had corresponded to point *g*, the system would behave in a different manner. If H_i is increased, the system will begin to follow the righthand branch of a hysteresis contour until the median curve *oac* is reached, after which it will follow this curve. On the other hand, if H_i is decreased the system will move toward the median curve cutting across hysteresis contours until the median curve is reached, after which it will follow downward along the median curve. If originally both B_i and H_i were positive and their values corresponded to point *d* on the right of the median curve, then an increase in H_i would cause the system to move along a hysteresis contour to the median line as shown. However,

decreasing H_i would cause the trace to cut across the hysteresis contours and the median line and then travel downward along the lefthand branch of the inner contour. Points such as e and f to the left of the median curve would behave in just the reverse manner.

Any time the system is in a state represented by a point off the median line, it is in a metastable condition, and if H_i is held constant and the sample is subjected to stress or temperature cycling, the system will move along a vertical line until it reaches the median curve.

Despite the apparent complexity of this behavior, it is possible to define certain properties which will serve to characterize it fairly completely. These are:

1. the saturation permeability,

$$ {}_s\mu_{ij} = \frac{\partial B_i}{\partial H_j}\Big)_{T,P,\ldots} \qquad \text{(when } H_j = {}_sH_j\text{)} $$

2. the zero point permeability,

$$ {}_0\mu_{ij} = \frac{\partial B_i}{\partial H_j}\Big)_{T,P,\ldots} \qquad \text{(when } B_i,\ H_j = 0\text{)} $$

3. the size of the hysteresis loop, described in terms of the ratio of the energy stored to the energy dissipated per cycle, when the system is cycled between positive and negative saturation values.

All of these are tensor quantities and vary with direction with respect to the crystallographic axes.

To summarize, the behavior of a system exhibiting extreme hysteresis is complicated by the fact that there is no unique functional relationship between the parameters except in the saturation range. Instead, the response of one parameter to a change in the other depends upon whether the parameter chosen as independent is increased or decreased, positive or negative, and upon its previous extreme value. The behavior of the system with respect to a particular pair of parameters can partially be characterized by specifying:

1. the value of the intensive parameter necessary to produce saturation,
2. the rate of change of the extensive parameter with respect to the intensive parameter at saturation,
3. the slope of the median curve at the point where both parameters are zero,
4. the area of the hysteresis loop when the intensive parameter varies between positive and negative saturation values,
5. the slope of the steep portion of the hysteresis curve, and
6. the value of the extensive parameter corresponding to the zero value of the intensive parameter.

6.6 Class IV: physically irreversible systems

We noted at the beginning of this chapter that one category of properties with which we would be concerned could be measured only by a test that would leave the test specimen permanently altered in some manner. We choose to refer to this as physical irreversibility, in contrast with the thermodynamic irreversibility that characterizes systems with hysteresis.

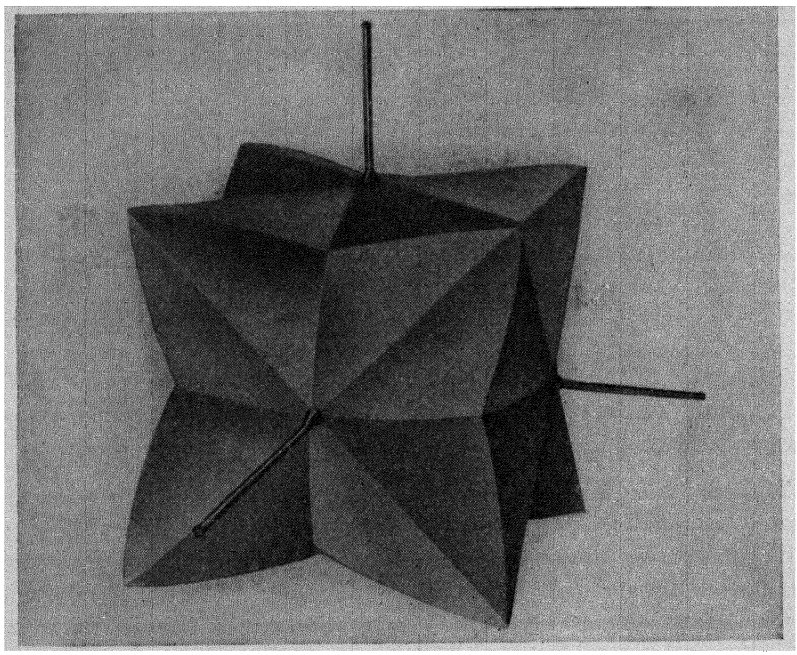


Figure 6.3 Directionality of ultimate tensile strength for aluminum. (From von Göler and Sachs, *Z. Techn. Physik* **8**, 586 (1927).)

Although of prime engineering importance, surprisingly, properties in this class do not have a precise quantitative meaning. They are only made quantitative by specifying a set of rather artificial test conditions. The results of these tests will have no absolute significance, but may serve to rank different materials with respect to the property tested. For example, there are at least two different criteria we might use to measure ductility, and three different criteria by means of which we might attach a number to the property called hardness. These provide different sets of numbers which will rank materials with respect to these qualities, but the different criteria may not even rank different materials in the same way.

Since the quantitative evaluation of a property in this class is only rela-

tive, it is not possible to extend the results of such tests into areas very much different from those under which the property is evaluated. For example, we may evaluate the ability of a material to carry a load by determining the axial stress required to cause mechanical failure of the material. However, this value cannot be used to predict the behavior of the material under a bi- or triaxial load in anything except a qualitative way.

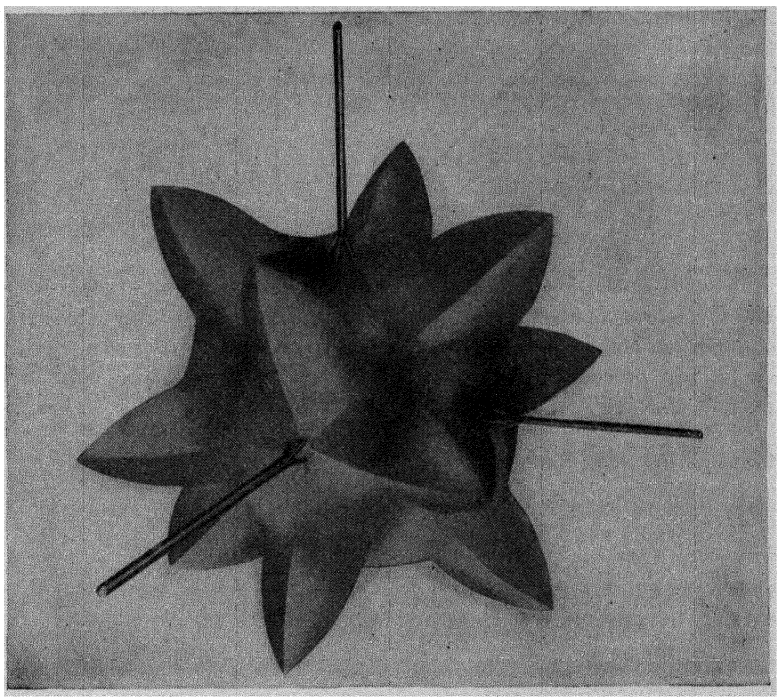


Figure 6.4 Directionality of elongation at break for aluminum. (From von Göler and Sachs, loc. cit.)

Even though the properties in this category are not defined in the precise mathematical sense that is possible for the first two categories and, to a lesser degree, the third category, it is still possible to correlate them with crystal symmetry when we are dealing with single crystal materials. Figures 6.3 and 6.4 show respectively the ultimate tensile strength and elongation at rupture for a single aluminum crystal. Along these same lines, the ease of lapping the surface of a diamond crystal depends strongly upon the density of atoms on the planes upon which the lapping occurs. Later, when we discuss the plastic properties of metals, we shall also show that the mobility of dislocations is strongly dependent upon the density of atoms on the plane through which the dislocation line is moving.

PROBLEMS

6.1 Consider a system in which stress and magnetic field intensity are the only extensive parameters that can be varied. Assume that these behave reversibly.

(a) Define four compliances and four rigidities that completely characterize the behavior of the system when any of the system parameters are varied.

(b) Which pair of compliances are equal? Which pair of rigidities? Develop the thermodynamic arguments that demonstrate this equality.

(c) Designate the rank of each tensor among the rigidities that were defined above.

(d) Derive the relationships between each of the compliances and all of the rigidities.

6.2 Given the symbolic equation

$$\alpha = \frac{-\lambda}{R_{TS}R_{\sigma\epsilon} - \lambda^2}$$

and the statement that λ is a rank two tensor, λ^2 is a rank four tensor whose components are $\lambda_{ij}\lambda_{kl}$, $\mathbf{R}_{\sigma\epsilon}$ a rank four tensor, and R_{TS} a scalar, outline the operational procedure for relating any particular component of α to the components of λ and $\mathbf{R}_{\sigma\epsilon}$. (*Hint:* see Sec. 3.4.)

6.3 Consider the set of compliances

$$\frac{\partial S}{\partial T}\Big)_H \qquad \frac{\partial S}{\partial \mathbf{H}}\Big)_T \qquad \frac{\partial \mathbf{B}}{\partial T}\Big)_H \qquad \frac{\partial \mathbf{B}}{\partial \mathbf{H}}\Big)_T$$

which characterize the behavior of a reversible system in which $\mathbf{H}$ and T are the only intensive parameters of importance.

(a) Write in extenso the components of the tensors $\partial S/\partial \mathbf{H}\big)_T$ and $\partial \mathbf{B}/\partial \mathbf{H}\big)_T$.

(b) Compare the corresponding components of the compliance $\partial \mathbf{B}/\partial \mathbf{H}\big)_T$, the rigidity $\partial \mathbf{H}/\partial \mathbf{B}\big)_S$, and the quantity $\partial \mathbf{H}/\partial \mathbf{B}\big)_T$.

6.4 Taking cognizance of the fact that certain of the quantities in the symbolic equation

$$R_{T,S} = \frac{C_{B,H}}{C_{S,T}C_{B,H} - C_{S,H}^2}$$

are tensors of different ranks, rewrite the equation in proper tensor form.

6.5 Let $\mathbf{A}$ be some vector extensive parameter functionally related to a rank two intensive tensor parameter $\mathbf{B}$ through the rank three compliance $\mathbf{C}$ by the equation

$$\mathbf{A} = \mathbf{C} : \mathbf{B}$$

The components of **A** and **B** can also be related through the set of equations

$$A_1 = d_{11}B_{11} + d_{12}B_{22} + d_{13}B_{33} + d_{14}B_{23} + d_{15}B_{13} + d_{16}B_{12}$$
$$A_2 = d_{21}B_{11} + d_{22}B_{22} + d_{23}B_{33} + d_{24}B_{23} + d_{25}B_{13} + d_{26}B_{12}$$
$$A_3 = d_{31}B_{11} + d_{32}B_{22} + d_{33}B_{33} + d_{34}B_{23} + d_{35}B_{13} + d_{36}B_{12}$$

Derive the relationships between the coefficients d_{ij} and the components of the **C** tensor.

6.6 The components of a vector extensive parameter **A** and a complementary vector intensive parameter **B** are functionally related, but the relationship is nonlinear. If

$$\mathbf{C}_{A,B} = \frac{\partial_{\mathbf{A}}}{\partial_{\mathbf{B}}}\Big)_{\text{all other intensives conts.,}}$$

develop the equation

$$\Delta\mathbf{A} = \mathbf{C}_{A,B} \cdot \Delta\mathbf{B} + \frac{1/2\partial\mathbf{C}_{A,B}}{\partial\mathbf{B}} : (\Delta\mathbf{B})^2 \ldots$$

What is the rank of the tensor $\partial\mathbf{C}_{A,B}/\partial\mathbf{B}$? Identify the mathematical nature of $(\Delta\mathbf{B})^2$.

6.7 Let ${}_e\mathbf{\varrho}$, ${}_t\mathbf{\varrho}$ and ${}_{et}\mathbf{\varrho}$ respectively represent electrical, thermal, and thermoelectric cross resistivities. Let ${}_e\mathbf{C}$, ${}_t\mathbf{C}$, and ${}_{et}\mathbf{C}$ respectively represent electrical, thermal, and thermoelectric cross conductivities.

(a) If the material is isotropic the ρ's and C's become scalars. In this case, derive the relationship

$${}_e\rho = \frac{{}_tC}{{}_tC{}_eC - {}_{et}C^2}$$

Note that if the cross effects are zero this reduces to ${}_e\rho = 1/{}_eC$.

(b) Show that if the cross effects are zero, but the resistivities and conductivities are tensors, ${}_e\mathbf{\rho} : {}_e\mathbf{C} = \mathbf{\delta}$.

(c) In the most general case, in which all quantities are tensors and the cross effects are not zero, show that either of the equations

$${}_e\mathbf{\varrho} = \frac{{}_t\mathbf{C}}{{}_t\mathbf{C} : {}_e\mathbf{C} - {}_{et}\mathbf{C} : {}_{et}\mathbf{C}}$$

or

$${}_e\mathbf{\varrho} : [{}_t\mathbf{C}{}_e\mathbf{C} - {}_{et}\mathbf{C}{}_{et}\mathbf{C}] = {}_t\mathbf{C}$$

are valid.

7

STRUCTURAL SYMMETRY AND NEUMANN'S PRINCIPLE

The properties defined in the first classes in Chap. 6 were generally tensor quantities. By regarding them in this manner, we may treat the anisotropy of solid materials in a natural straightforward way. The structural symmetry of a solid will impose certain limitations upon the relationships between different components of the same tensor property: this is the major point we shall explore in this chapter. As a preliminary, we must first consider the symmetry of crystals and the methods used to characterize this symmetry.

7.1 Crystal classes and symmetry

We may subdivide the six crystal systems into thirty-two classes according to their symmetry. Table 7.1 lists these classes along with the minimum elements of symmetry that are required to qualify a crystal for membership in a class. The classes are arranged in order of increasing symmetry, and membership in a class requires that it have at least the minimum elements of symmetry specified for the class and be lacking in some of the symmetry elements of the higher class.

The symmetry elements of a crystal are defined by a set of transformation operations which leave the basic cell in a state indistinguishable from its original one. These operations consist of rotations about certain axes, reflections through planes or points, and combinations of these. Such axes,

Table 7.1 *Crystal classes*

Crystal system	Crystal class	Minimum symmetry elements
I. Triclinic	1. Triclinic-pedial	1
	2. Triclinic-pinacoidal	$\bar{1}$
II. Monoclinic	3. Monoclinic-domatic	m
	4. Monoclinic-sphenoidal	2
	5. Monoclinic-prismatic	$2/m$
III. Orthorhombic	6. Orthorhombic-pyramidal	mm
	7. Orthorhombic-disphenoidal	222
	8. Orthorhombic-dipyramidal	mmm
IV. Tetragonal	9. Tetragonal-disphenoidal	$\bar{4}$
	10. Tetragonal-pyramidal	4
	11. Tetragonal-dipyramidal	$4/m$
	12. Tetragonal-scalenohedric	$42m$
	13. Ditetragonal-pyramidal	$4mm$
	14. Tetragonal-trapezohedric	42
	15. Ditetragonal-dipyramidal	$4/mmm$
V. Trigonal-hexagonal	16. Trigonal-pyramidal	3
	17. Rhombohedric	$\bar{3}$
	18. Ditrigonal-pyramidal	$3m$
	19. Trapezohedric	32
	20. Ditrigonal-scalenohedric	$\bar{3}m$
	21. Hexagonal-pyramidal	6
	22. Trigonal-dipyramidal	$\bar{6}$
	23. Hexagonal-dipyramidal	$6/m$
	24. Ditrigonal-dipyramidal	$\bar{6}m2$ or $\bar{6}2m$
	25. Dihexagonal-pyramidal	$6mm$
	26. Hexagonal-trapezohedric	62
	27. Dihexagonal-dipyramidal	$6/mmm$
VI. Cubic	28. Tetrahedric-pentagon-dodecahedric	23
	29. Dyakis-dodecahedric	$m3$
	30. Hexakis-tetrahedric	$43m$
	31. Pentagon-ikositetrahedric	43
	32. Hexakis-octahedric	$m3m$

planes, and points are called, respectively, *axes of symmetry*, *planes of symmetry*, and *centers of symmetry*. These operations are defined in Table 7.2.

By way of explanation of Table 7.2, if successive rotations through 120° about some axis would produce self-coincidence of the basic cell, that axis would be a 3-fold axis of symmetry. The combinations of elements $44m$ states that two 4-fold axes of symmetry and a reflection plane are present with one of the axes lying in the reflection plane. In contrast, $44/m$ specifies

the same symmetry elements except that one of the axes of symmetry is perpendicular to the plane of symmetry. Some combinations of symmetry elements necessarily imply the existence of others. For example, *mm* also implies the presence of a 2-fold axis of symmetry. Most of the members of a

Table 7.2 *Symmetry transformations*

Operation producing self-congruence	Name	Symbol
Rotation about some axis through an angle $360°/n$	n-Fold axis of symmetry	n
Reflection through a plane	Plane of symmetry	m
Reflection through a point	Center of symmetry	$\bar{1}$
Rotation about an axis through $360°/n$ followed by reflection through a point on the axis of rotation	Rotation-inversion axis	$\bar{n}$
	Axis of symmetry perpendicular to a plane of symmetry	n/m
	Axis of symmetry lying in a plane of symmetry	nm

class will have other symmetry elements in addition to those which qualify them for the class.

Multiple planes of symmetry require some special comment. If two or more planes of symmetry intersect along a line, they will all make equal angles with each other. For example, the designation *mmm* could imply that there are three planes of symmetry intersecting along a line all having dihedral angles of 60° with one another. Alternatively the designation could mean that there are three mutually perpendicular planes of symmetry.

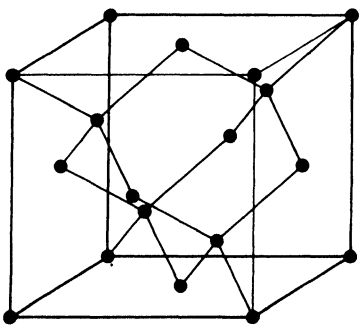


Figure 7.1 Diamond lattice.

The application of these symmetry elements may be illustrated by classifying the diamond lattice of Fig. 7.1. Instead of attempting to find all the symmetry elements of the basic cell of the diamond lattice, we shall follow a cut-and-try procedure until we have found the proper classification. All the members of the cubic system will possess at least one 3-fold axis of symmetry lying along a diagonal of the cube. Beginning with class 28, whose minimum symmetry elements are 23, we find that in addition to a 3-fold axis of symmetry the diamond lattice also has a 2-fold axis perpendicular to each face. Therefore, the diamond lattice could be in this class provided it is lacking in the symmetry elements of higher classes. Class 29 has the symmetry elements $m3$. The diamond basic cell does have planes of symmetry that intersect each face along a face diagonal and include the 3-fold axis. This consideration

removes the diamond lattice from class 28 and qualifies it as a possible member of class 29. The symmetry elements of class 30 are $\bar{4}3m$. We note that each 3-fold axis of symmetry of the diamond lattice lies in a plane of symmetry, hence the lattice has the symmetry element $3m$. It also has a 4-fold rotation-inversion axis perpendicular to each face of the basic cell. This factor removes it from class 29, and we place it tentatively in class 30. It is lacking in a 4-fold rotation axis. This eliminates the structure from class 31, leaving class 30 as the one to which the diamond lattice must belong.

When a crystal is subjected to one of the symmetry operations described, each atom of the crystal is mentally moved from its original position, but upon completion of the operation, every point that was initially occupied by an atom of a particular kind will still be occupied by an atom of the same kind. Hence the final state of the crystal after the symmetry transformation will be indistinguishable from its original state. Instead of imagining that this transformation is performed upon the crystal itself, we might choose some Cartesian coordinate system x_i and locate every point of the crystal in this coordinate system. Then a symmetry operation valid for the crystal under consideration may be applied to the coordinate system to transform it into a new set of coordinates such that a particular set of coordinate numbers locating an atom in the original coordinate system will locate an atom of the same kind in the new coordinate system. To clarify this idea, consider Fig. 7.2. This shows a simple array of two different kinds of object. One of these is represented by a solid circle and the other by an open circle. In the coordinate system x_i, the coordinates of the open circles are (3, 0, 0) and (−3, 0, 0), whereas those of the solid circles are (0, 2, 0) and (0, −2, 0). If we generate a new coordinate system y_i by rotation of the x_i system through 180° about the x_3 axis, then the same points, (3, 0, 0) and (−3, 0, 0), in the y_i coordinate system will be occupied by open circles and similarly the points (0, 2, 0) and (0, −2, 0) will be occupied by solid circles. When using the crystal symmetry operations, we moved the individual points about in such a way that the array was unchanged; here we are transforming a coordinate system so that kinds of atoms occupying particular points in either coordinate system are unchanged.

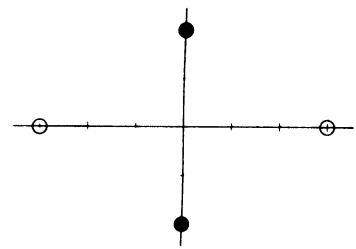


Figure 7.2 A transformation of coordinates corresponding to rotation through 180° about an axis perpendicular to the plane of the paper would produce a new coordinate system such that a set of numbers locating either a closed or an open circle would be valid in either coordinate system.

We have already seen that transformation matrices provide a convenient method of transforming from one coordinate system to another, and each symmetry operation may be accomplished by the application of a set of one

or more transformation matrices. Table 7.3 lists a set of transformation matrices corresponding to the symmetry operators defined previously. Symmetry operators such as a 3-fold axis of symmetry require that the crystal come into self-coincidence under a rotation of either 120° or 240°. Thus the operation could be described by either of two different transformation

Table 7.3

Symmetry	Orientation of coordinates	Transformation matrices
$\bar{1}$	Center of symmetry located at the origin of the coordinate system	$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$
2	Twofold axis of rotation with the axis of symmetry coincident with the x_1 axis	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$
m	Plane of symmetry with the x_2 and x_3 axes lying in the plane of symmetry	$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
3	Three-fold axis of symmetry with the axis of symmetry coincident with the x_1 axis (rotation through 120°)	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{2} & \sqrt{\frac{3}{2}} \\ 0 & -\sqrt{\frac{3}{2}} & -\frac{1}{2} \end{bmatrix}$
4	Fourfold axis of symmetry with x_1 axis of coincident with the axis of symmetry (rotation thru 90°)	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}$
$2m$	Twofold axis of symmetry, axis of symmetry lying in the reflection plane and coincident with the x_1 axis and the x_2 axis lying in the reflection plane	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
$2/m$	Twofold axis of symmetry perpendicular to a reflection plane with the axis of symmetry coincident with the x_1 axis	$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$

matrices. In the applications we shall meet, it will usually suffice to apply one of these, and it is immaterial which is used, except as may be indicated by algebraic convenience. Table 7.3 lists only one transformation matrix for these cases. Other symmetry operations such as $2/m$ correspond to a transformation matrix derived from the product of two different matrices. In this case it would be the product of the transformation matrix corresponding to a rotation through 180° and the matrix corresponding to reflection through a plane.

7.2 Neumann's principle

The concept upon which the subsequent developments of this chapter rest is very simple but has extensive consequences. It arises very naturally from the symmetry transformation considered in the last section. Just as the symmetry transformation left the structure in a state indistinguishable from its original one, so it will leave the components of a tensor representing properties associated with that structure. To phrase it more formally, Neumann's principle states that the *components* of a tensor representing a property will remain invariant under a transformation of coordinates governed by a symmetry operation valid for the structure under consideration. The reader should not confuse this with the invariance of the tensor itself, which holds under any orthogonal transformation of coordinates. Even though the tensor remains invariant, its components ordinarily do change according to a specific formula. Here the symmetry of the structure imposes constraints such that the components will not change under certain transformations.

7.3 Properties of materials represented by tensors of rank one

It has been widely noted that crystals having a center of symmetry are never pyroelectric or pyromagnetic. The pyroelectric and pyromagnetic coefficients are vector properties of a material, and their absence in crystals having a center of symmetry is a logical consequence of the following arguments.

Let $\mathbf{V}$ be a vector property of a crystal with components v_1, v_2, and v_3 in an arbitrary coordinate system. If the crystal has a center of symmetry, then by Neumann's principle the components of the vector must remain invariant under a transformation of coordinates corresponding to reflection through a point. Such a transformation may be accomplished by the matrix equation

$$\mathbf{V}' = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \mathbf{V}$$

From the transformation we obtain

$$v_1' = -v_1, \qquad v_2' = -v_2, \qquad v_3' = -v_3$$

but the application of Neumann's principle requires that

$$v_1' = v_1, \qquad v_2' = v_2, \qquad v_3' = v_3$$

The inescapable conclusion is that all of the components of the vector must be zero since they are all equal to their own negative. Thus, a crystal having a center of symmetry will be lacking in any vector property.

Using a similar line of reasoning, we may deduce that if a crystal does have one or more vector properties and also has an axis of symmetry, the vector property will be parallel to the axis of symmetry. We may also deduce that two different axes of symmetry will cause the symmetry of the crystal to be too high to allow it to have any vector properties.

7.4 Properties that may be represented by tensors of rank two

Typical of the properties in this class are the thermal expansion, the magnetic permeability, the dielectric permittivity, the thermal conductivity, and the electrical conductivity coefficients. The arguments presented here are valid for all of these since they depend only upon the rank of the tensor required to describe the properties.

Before exploring any of these arguments in detail, let us first recall that rank two tensor properties are always symmetric. In an arbitrary coordinate system, there will be a total of six coefficients describing the phenomenon, but these are not all independent. In a principal coordinate system there will be three nonzero components and, consequently, each of the properties in this category may be characterized by a maximum of three independent coefficients.

By considering the effect of the symmetry operator describing the least symmetric class of each of the crystal systems, we can best show how the characteristics of a rank two tensor property change with progressively increasing symmetry. In the case of the most asymmetric type of crystal with only a onefold axis, there would be no restrictions placed upon the tensor properties by symmetry. Therefore, the orientation of the principal axes of each tensor property with respect to the crystallographic axes could not be deduced from symmetry considerations. Although it is not self-evident, neither does the existence of a center of symmetry place any restrictions upon tensor properties of rank two. The other classes we will take up are monoclinic-domatic with the symmetry element m, orthorhombic-pyramidal with the symmetry elements mm, tetragonal-disphenoidal with the symmetry $\bar{4}$. It turns out that the symmetry elements of the classes of higher symmetry impose no further conditions upon the characteristics of the tensor until class 28, whose name is too long to bear repetition, with symmetry 23. In each case, the three points that we shall examine are the orientation of the principal axes with respect to the crystallographic axes, the number of independent constants required to characterize the tensor, and whether or not the property ever becomes isotropic on a plane or in space.

The monoclinic-domatic class. In the monoclinic system, one of the crystallographic axes is perpendicular to the other two, while the third angle is

arbitrary. The plane of symmetry characteristic of this class will be perpendicular to the b axis using the system designating axes that was described in Table 4.1. We shall represent the components of a rank two tensor property in a coordinate system so oriented that the x_i axis is perpendicular to the plane of symmetry and the other two axes lie in the plane of symmetry. Then the components of the tensor must be invariant under the transformation of coordinates governed by the equation

$$(y_i) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} (x_i)$$

Using the general transformation formula for tensors of rank two

$$T'_{ij} = a_{ik}a_{jl}T_{kl}$$

where the a's are the appropriate elements of the transformation matrix, we may compute

$$T'_{11} = a_{11}a_{11}T_{11} + a_{11}a_{12}T_{12} + a_{11}a_{13}T_{13} + a_{12}a_{12}T_{22} + a_{12}a_{11}T_{21} + a_{12}a_{13}T_{23} + a_{13}a_{11}T_{31} + a_{13}a_{12}T_{32} + a_{13}a_{13}T_{33}$$

In this case all of the terms are zero except the first, and the equation reduces to

$$T'_{11} = T_{11}$$

In a similar manner we may compute

$$T'_{22} = T_{22}, \quad T'_{12} = -T_{12}, \quad T'_{33} = T_{33}, \quad T'_{13} = -T_{13}, \quad T'_{23} = T_{23}$$

Since Neumann's principle requires the primed quantities to be equal to the unprimed quantities, T_{12} and T_{13} must be zero because they are equal to their own negatives.

A rotation about the x_1 axis is capable of reducing T_{23} to zero, in which case the coordinates will be a set of principal coordinates. We may show this by generating a coordinate system z_i by performing the transformation

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{bmatrix}$$

upon the x_i coordinate system. Using a double prime to indicate the components of the tensor in the z_i system

$$T''_{23} = -\sin\theta\cos\theta T_{22} + \sin\theta\cos\theta T_{33} + \cos^2\theta T_{23}$$

We may obtain unique solutions for the angle θ with arbitrary values of T_{22}, T_{33}, and T_{23} when we set T''_{23} equal to zero. Consequently, the set z_i will be principal axes. If any arbitrary value of θ would have satisfied the previous

equation, then we would have concluded that these properties were isotropic on the plane of symmetry. However, since only unique values of θ do so, we conclude that this does not occur.

To summarize, the rank two tensor properties of crystals in the monoclinic-domatic class will have three independent coefficients. There is no plane upon which the properties become isotropic, and the only conclusion that we may reach about the orientation of the principal axes is that one of these will be parallel to the b axis. (See Fig. 7.3.) These same considerations apply to the entire monoclinic system.

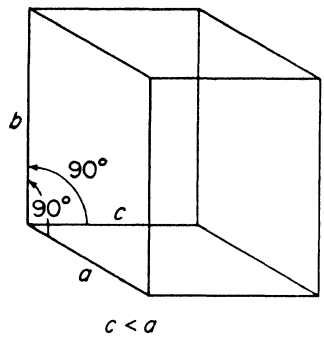


Figure 7.3 Basic cell of a monoclinic crystal.

The orthorhombic-pyramidal class. In orthorhombic crystals, the crystallographic axes will coincide with a set of principal axes of any rank two tensor property. The two reflection planes characterizing this class are mutually perpendicular and intersect along a line parallel to one of the crystallographic axes. Consequently, the other two axes will each be parallel to a reflection plane. We choose our coordinate system with the x_2 axis coincident with the line of intersection of the two reflection planes and with the x_1 axis lying in plane number 1 and the x_3 axis lying in plane number 2. We have already shown that the presence of a reflection plane number 2 will cause the components T_{12} and T_{13} to be zero in a coordinate system chosen in this manner. By Neumann's principle, the components of the tensor must also be invariant under a transformation of coordinates corresponding to reflection through plane number 1. This transformation will be governed by the matrix equation

$$(z_i) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} (x_i)$$

Upon applying the transformation formula to the components of the tensor, we find the relationships

$$T'_{11} = T_{11}, \quad T'_{22} = T_{22}, \quad T'_{33} = T_{33}, \quad T'_{23} = -T_{23}$$

Consequently T_{23} must be zero, and we have proven our initial assertion: that a set of coordinate axes coincident with the crystallographic axes will be principal axes. This class of crystals is anisotropic on all planes.

The tetragonal-disphenoidal class. The transformation matrix associated with the fourfold rotation-inversion axis characterizing the symmetry of class

9 crystals may be derived by multiplying the matrix operators for rotation about an axis through 90° and for reflection through the origin. Again we choose our x_1 axis coincident with the axis of symmetry. Then the appropriate transformation matrix is obtained from

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}$$

where the first two matrices are obtained from Table 7.3.

Applying the transformation formula to the components of the tensor gives the set of relationships

$$T'_{11} = T_{11}, \qquad T'_{22} = T_{33}, \qquad T'_{33} = T_{22}$$
$$T'_{23} = -T_{23}, \qquad T'_{12} = -T_{13}, \qquad T'_{13} = -T_{12}$$

When we apply Neumann's principle to this, we must conclude that the following relationships hold

$$T_{22} = T_{33} \qquad \text{and} \qquad T_{12} = -T_{13}$$

Thus far we have only imposed the conditions arising from a 90° rotation about the x_1 axis followed by a reflection through the origin. The tensor components must also be invariant under a rotation of 180° followed by a reflection through the origin. This is equivalent to the operation associated with a reflection plane containing the x_2 and x_3 axes. We have already shown that this will cause T_{12} to be zero.

In summary, we have $T_{22} = T_{33}$, $T_{12} = 0 = -T_{13}$, and $T_{23} = -T'_{23} = 0$. Consequently, the rank two tensor properties of crystals having this or higher symmetry (but less than that of class 28) will have only two independent coefficients.

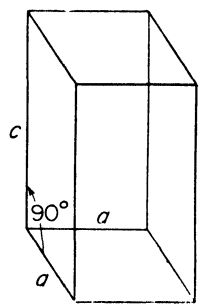


Figure 7.4 Outline of polyhedron associated with a basic cell belonging to the tetragonal system.

Since initially we only assumed that the coordinate system chosen had its x_1 axis coincident with the axis of symmetry, we may conclude that for this class all coordinate axes with one axis aligned with the axis of symmetry are principal axes. Therefore, these crystals will be isotropic on planes perpendicular to the axis of symmetry. The axis of symmetry is parallel to the c crystallographic axis of Fig. 7.4.

The tetrahedric-pentagon-dodecahedric class. Class 28 belongs to the cubic system, and the threefold axis of symmetry is directed along a body diagonal of the basic cell, while the twofold axis is directed perpendicularly to one of a pair of faces of the cube. The matrix associated with rotation

through 120° about the threefold axis may be derived by referring to Fig. 7.5. The coordinate system x_i is chosen coincident with the crystallographic axes, and the line aa' is the threefold axis of symmetry which makes the angle $\cos^{-1} 1/\sqrt{3}$ with each of the axes. A 120° rotation about this axis will

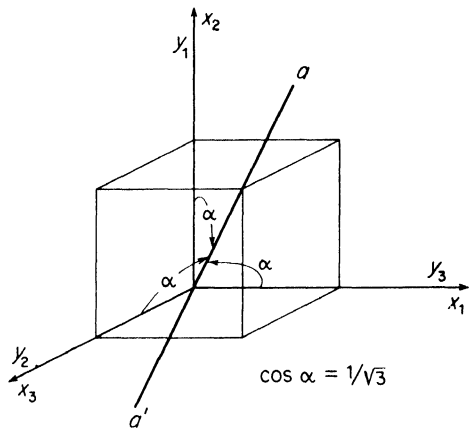


Figure 7.5 All of the basic cells of the cubic system have a set of 3-fold axes of symmetry along their body diagonals.

generate a coordinate system y_i which is related to the x_i system by the set

$$x_1 = y_2, \qquad x_2 = y_3, \qquad x_3 = y_1$$

Accordingly, the transformation matrix is

$$(x_i) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} (y_i)$$

Applying the transformation formula to the tensor produces the relationships

$$T'_{11} = T_{22}, \qquad T'_{22} = T_{33}, \qquad T'_{33} = T_{11}$$
$$T'_{12} = T'_{13} = T'_{23} = 0$$

Upon application of Neumann's principle, we find

$$T_{11} = T_{22} = T_{33}$$

The reader may easily prove to himself that the components of the tensor remain invariant under a rotation about any of the coordinate axes. If this is true, then they will remain invariant under any combination of these. Consequently, all axes will be principal axes, and materials having this or higher symmetry will be isotropic with respect to all of their rank two tensor properties, which may be considered to be scalar quantities. Occasionally an

author will leap to the conclusion that the properties of a cubic material will vary from a direction parallel to the edge of the cube to a direction parallel to a face or body diagonal. This may be true of rank three or four tensor properties, but it is simply not possible for rank two properties to be other than isotropic for structures having this degree of symmetry.

7.5 Convention for choosing the orientation axes

For crystals having a symmetry lower than that of the orthorhombic class, there is no way from symmetry considerations alone to specify the orientation of the principal axes with respect to the crystallographic axes. Experimentally it would be possible to make a measurement of any tensor property of these crystals using a coordinate system having an arbitrary but known orientation with respect to the crystallographic axes. Once all of the tensor components in this coordinate system were known from direct experiment, the orientation of the principal axes and the values of the tensor components in the principal coordinate system could be calculated. This reduction of data can be rather laborious and would be unnecessary if everyone would simply use the same coordinate system chosen by convention. This would entail reporting six tensor components in the conventional coordinate system rather than three components in the principal axes along with information about the orientation of these with respect to the crystallographic axes.

Such a convention has been adopted[1] and is now widely used. We did not follow this convention in the previous section because in attempting to grasp a new set of ideas there is often a danger in confusing that which is basic with that which is merely conventional. Now we must get in step with the rest of the scientific community, and this convention will be followed in the remainder of the book.

The convention requires, first of all, that the crystallographic axes be labeled in the same way that they are in Table 4.1. Then the x_3 axis is chosen to be coincident with the c-crystallographic axis, and the x_1 axis is chosen to lie in the plane of the a- and c-crystallographic axes. This conventional orientation of the coordinate system must be considered arbitrary, and any rank two tensor will have six distinct components in it. Most authors call the coefficients defined in this way "independent." Of course, they are not, since one can always reduce their number to three by a principal axis transformation. Table 7.4 lists the matrices of the tensor obtained by using this convention and shows how increasing symmetry causes progressively more coefficients to become zero. The results agree with the developments of Sec. 7.4. The difference in the subscripts of certain of the nonzero terms is due only to the fact

[1] Warren P. Mason, *Piezoelectric Crystals and Their Application to Ultrasonics* (New York: D. Van Nostrand Company, Inc., 1950).

Table 7.4 *Matrices of components of rank two tensor properties*

System	Matrix	System	Matrix
Triclinic	$\begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{12} & T_{22} & T_{23} \\ T_{13} & T_{23} & T_{33} \end{bmatrix}$	Monoclinic	$\begin{bmatrix} T_{11} & 0 & T_{13} \\ 0 & T_{22} & 0 \\ T_{13} & 0 & T_{33} \end{bmatrix}$
Orthorhombic	$\begin{bmatrix} T_{11} & 0 & 0 \\ 0 & T_{22} & 0 \\ 0 & 0 & T_{33} \end{bmatrix}$	Tetragonal and trigonal-hexagonal	$\begin{bmatrix} T_{11} & 0 & 0 \\ 0 & T_{11} & 0 \\ 0 & 0 & T_{33} \end{bmatrix}$
Cubic	$\begin{bmatrix} T_{11} & 0 & 0 \\ 0 & T_{11} & 0 \\ 0 & 0 & T_{11} \end{bmatrix}$		

that we did not follow the convention in choosing the orientation of our coordinate system.

7.6 Rank three tensor properties

The most important rank three tensor properties of a material are those relating electric field intensity to strain and dielectric displacement to stress. These characterize the phenomenon of piezoelectricity, and will be treated in some detail in Chap. 12.

Any tensor of rank three will have twenty-seven components. However, a rank three tensor property must reduce to a vector by contraction with a rank two tensor. Since a symmetric rank two tensor has only six distinct components, a rank three tensor property will, therefore, have $3 \times 6 = 18$ distinct components. These could be reduced to nine independent coefficients by using a coordinate system in which the contracted tensor was diagonal, but ordinarily the conventional coordinate system is used, in which case 18 of the coefficients are considered independent. In this coordinate system, the terms $T_{ijk} = T_{ikj}$. Instead of listing all 27 components, normally only the eighteen components

$$\begin{bmatrix} T_{111} & T_{122} & T_{133} & T_{123} & T_{113} & T_{112} \\ T_{211} & T_{222} & T_{233} & T_{223} & T_{213} & T_{212} \\ T_{311} & T_{322} & T_{333} & T_{323} & T_{313} & T_{312} \end{bmatrix}$$

are listed in the form of a matrix, where it is understood that T_{112} equals T_{121}, T_{223} equals T_{232}, etc.

Using the transformations characteristic of the symmetry of each of the crystal classes and applying Neumann's principle, we may show how symmetry causes certain of the coefficients to become zero in the same manner as in Sec. 7.4. In the case of rank three properties we cannot generalize for an

entire crystal system, but instead each class must be treated separately. The results of such a set of calculations are tabulated in Appendix A. It would be needlessly profligate of time and space to show all of the calculations in detail, but we shall prove that all of the coefficients are zero when a center of symmetry is present, and we shall pick one other example to show the details of proving which terms are zero.

Center of symmetry. The transformation matrix corresponding to the symmetry operator $\bar{1}$ is

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

For tensors of rank three, the general transformation formula is

$$T'_{ijk} = a_{il}a_{jm}a_{kn}T_{lmn}$$

and by Neumann's principle

$$T'_{ijk} = T_{ijk}$$

Now the set of coefficients a_{il}, a_{jm}, a_{kn} will be zero unless the following relationships hold among the subscripts

$$i = l, \qquad j = m, \qquad k = n$$

in which case the set is equal to -1. Consequently, the transformation formula reduces to

$$T'_{ijk} = -T_{ijk}$$

and, from Neumann's principle, we may therefore conclude

$$T'_{ijk} = T_{ijk} = -T_{ijk} = 0$$

In other words, crystals having a center of symmetry cannot have any rank three tensor properties.

Orthorhombic-disphenoidal (class 7). This example was picked at random to illustrate a case in which not all of the coefficients are zero. The symmetry operators characterizing this class are 222; the axes of symmetry are parallel to the crystallographic axes which are coincident with the conventional coordinate axes.

The transformation matrices corresponding to three mutually perpendicular twofold axes of symmetry which are coincident with the coordinate axes are

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

By Neumann's principle, the tensor components must remain invariant under successive applications of these transformations. This and the usual transformation formula are

$$T'_{ijk} = a_{il}a_{jm}a_{kn}T_{lmn} \qquad \text{and} \qquad T'_{ijk} = T_{ijk}$$

With a little care, we can kill off two birds with one cliché here. It doesn't matter in which order the transformation matrices are imposed. Notice that the symmetry of the monoclinic-sphenoidal class is 2 with the axis of symmetry parallel to the x_3 axis. We have to impose this symmetry condition anyway and by imposing it first, we may compute the matrix of the rank three tensor components for monoclinic-sphenoidal crystals as an incidental part of our calculation of the matrix of class 7 crystals.

In this case, we can tell from the transformation matrix that the set of coefficients, a_{il}, a_{jm}, a_{kn} will be zero unless

$$i = l, \qquad j = m, \qquad k = n$$

and will be equal to $+1$ whenever the subscripts consist of a combination of 1, 2, and 3, two 2's and a 3, two 1's and a 3, or all 3's. Any term having other than these combinations of subscripts will necessarily be zero, and the matrix of the coefficients of a rank three tensor property of a monoclinic sphenoidal crystal reduces to

$$\begin{bmatrix} 0 & 0 & 0 & T_{123} & T_{113} & 0 \\ 0 & 0 & 0 & T_{223} & T_{213} & 0 \\ T_{311} & T_{322} & T_{333} & 0 & 0 & T_{312} \end{bmatrix}$$

Upon this set of tensor components we impose the further condition that they be invariant under the transformation of coordinates effected by the matrix

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

Again the transformation is governed by the equation

$$T'_{ijk} = a_{il}a_{jm}a_{kn}T_{lmn}$$

with Neumann's principle embodied in the equation

$$T'_{ijk} = T_{ijk}$$

The set of coefficients a_{il}, a_{jm}, a_{kn} will be zero unless

$$i = l, \qquad j = m, \qquad k = n$$

and the righthand side of the tensor transformation formula will consist of

only the term

$$T'_{ijk} = a_{i(i)}a_{j(j)}a_{k(k)}T_{ijk}$$

Thus, whenever the set of a's is equal to -1, this will reduce to

$$T'_{ijk} = -T_{ijk} = T_{ijk} = 0$$

The set of a's will be negative for any combination of subscripts involving two 1's and a 3, two 2's and a 1, two 3's and a 1, two 2's and a 3, three 1's, or three 3's.

Applying these restrictions to the tensor components obtained from the first rotation, we have, for the matrix of the tensor components

$$\begin{bmatrix} 0 & 0 & 0 & T_{123} & 0 & 0 \\ 0 & 0 & 0 & 0 & T_{213} & 0 \\ 0 & 0 & 0 & 0 & 0 & T_{312} \end{bmatrix}$$

Proceeding in the same manner and transforming with the matrix operator

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

we find that no more terms become zero, and that the nonzero components of rank three tensor properties of a crystal of this class are

$$T_{123} = T_{132}, \qquad T_{213} = T_{231}, \qquad T_{312} = T_{321}$$

Isotropy of rank three tensor properties. Perhaps the most efficient way to decide whether any of the members of the various crystal classes is isotropic in space or on a plane would be to first impose the conditions for isotropy upon a general tensor, then compare the matrices of the resulting tensors with those of Table 7.4.

We could impose the conditions for isotropy on a plane perpendicular to the x_3 axis by requiring invariance of the components under the following transformation of coordinates

$$(y_i) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} (x_i)$$

Then requiring invariance of the components under an arbitrary rotation around either of the axes would assure complete isotropy. However, this headlong approach would involve the computation of 18 components, each of which might have 18 terms. Instead, we note that the matrix of the properties tensor of class 26 (Appendix A) was deduced using a sixfold axis of symmetry parallel to the z axis. If we begin with the tensor in this form, then

test the effect of an additional arbitrary rotation, we can reach the same conclusion with less effort and no loss of generality.

The 18 components of this tensor can be arranged in the matrix form

$$\begin{bmatrix} 0 & 0 & 0 & T_{123} & T_{113} & 0 \\ 0 & 0 & 0 & T_{223} & T_{213} & 0 \\ T_{311} & T_{322} & T_{333} & 0 & 0 & 0 \end{bmatrix}$$

with $T_{113} = T_{223}$, $T_{311} = T_{322}$, and $T_{123} = -T_{213}$. Under the transformation indicated previously, we may write

$$T_{113} = T'_{113} = -\cos\theta \sin\theta T_{123} + \cos^2\theta T_{113} + \sin^2\theta T_{223} - \sin\theta \cos\theta T_{213}$$

where the zero terms due to the occurrence of a_{13}, a_{23}, a_{31}, and a_{32} have been omitted. Since $T_{123} = -T_{213}$ and $T_{113} = T_{223}$, this reduces to

$$T'_{113} = T_{113}$$

In a similar manner, we find that all of the other terms remain invariant under an arbitrary rotation about the z axis. Thus, in order to be isotropic on planes perpendicular to the z axis, the matrix of the components of a rank three tensor property will be in the form

$$\begin{bmatrix} 0 & 0 & 0 & T_{123} & T_{113} & 0 \\ 0 & 0 & 0 & T_{113} & -T_{123} & 0 \\ T_{311} & T_{311} & T_{333} & 0 & 0 & 0 \end{bmatrix}$$

The crystal classes possessing a transverse isotropy are 10, 12, 14, 23, 24, and 25.

If we now attempt to rotate the coordinate system about the x_1, we find that the components are functions of the angles, and we must conclude that none of the crystal classes is isotropic in space with respect to a rank three tensor property.

7.7 Rank four tensor properties

The elastic rigidities and compliances are the only members of this class. Some of the characteristics of these tensors were examined in Chap. 3, and the next chapter is devoted exclusively to elasticity. Therefore, it seems reasonable to defer a discussion of rank four tensor properties until then.

PROBLEMS

7.1 (a) Classify the rutile structure (p. 55) according to its symmetry.

(b) A bcc structure is distorted as follows: one of the cubic axes is elongated; the other two are contracted by equal amounts. All axes are still mutually

perpendicular. The central atom is moved off the point of intersection of the body diagonals in a direction parallel to the elongated axis. All of the atoms are identical. Classify the new structure according to its symmetry.

(c) The zincblende structure consists of an fcc array of zinc atoms with sulfur atoms located at the points $(\frac{1}{4}\ \frac{1}{4}\ \frac{1}{4})$, $(\frac{1}{4}\ \frac{3}{4}\ \frac{3}{4})$, $(\frac{3}{4}\ \frac{1}{4}\ \frac{3}{4})$, $(\frac{3}{4}\ \frac{3}{4}\ \frac{1}{4})$. This differs from the diamond structure only to the degree that the atoms internal to the basic cell are different from those located on the surface of the basic cell. Place this structure in its proper crystal class.

7.2 Prove that if a structure has two mutually perpendicular two fold axes of symmetry, it will necessarily have three mutually perpendicular axes of symmetry.

7.3 Prove that three mutually perpendicular planes of symmetry imply that each plane of symmetry is perpendicular to a twofold axis of symmetry.

7.4 Taking the x_2 coordinate axis as the axis of symmetry, find the transformation matrix corresponding to the cases where the axis of symmetry is threefold, fourfold, and sixfold.

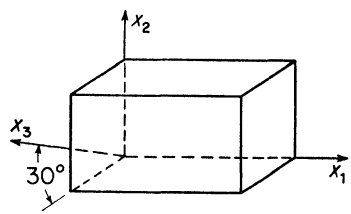
7.5 Verify the data in Appendix A concerning the rank three tensor properties of crystal class 10.

7.6 A single crystal of class 7 has the external shape of a rectangular parallelopiped but the crystallographic axes do not coincide in direction with the external edges of the crystal. Discuss, qualitatively, the general nature of the change in shape of the crystal that might be expected when it is heated.

7.7 With reference to Prob. 6, the components of the thermal expansion tensor in a set of principal axes are

$$\alpha_{11} = 2 \times 10^{-6}\ \text{deg C}^{-1} \quad \alpha_{22} = 3 \times 10^{-6}\ \text{deg C}^{-1} \quad \alpha_{33} = 5 \times 10^{-6}\ \text{deg C}^{-1}$$

The orientation of the crystallographic axes with respect to the geometric axes of the crystal are shown below



Calculate the strain produced by 100°C. temperature rise. Express this in terms of its components in a set of coordinates coincident with the edges of crystal.

7.8 Show that, in an orthorhombic crystal, the electric current density vector and the electric field intensity vector will only coincide in direction when they are both coincident with a crystallographic axis.

7.9 Outline the steps that would be necessary to characterize the magnetic permeability (this is a rank two tensor property) of a metal belonging to crystal class 27.

7.10 Demonstrate that if a crystal of class 10 is pyroelectric, the pyroelectric axis will coincide with the c-crystallographic axis (see Table 4.1).

7.11 In a recent letter to the editor of the *Journal of Applied Physics*, an investigator reported upon the results of thermal expansion measurements on silicon single crystals (diamond structure). He found a slight difference between measurements in the $\langle 100 \rangle$ and the $\langle 111 \rangle$ directions, but admitted that the difference was within the limits of his experimental error. Criticize his approach to the problem.

8

ELASTICITY AND PLASTICITY

8.1 Strain

Stress and strain are central concepts of any treatment of elasticity. Stress was explained in Chap. 3 and requires no further comment. Strain, however, is basically more complex and requires our critical attention here.

If we distort a continuous medium, we will cause a change in the location of the points of the medium relative to one another. These displacements may be used to characterize the distortion in the following way. Figure 8.1 shows a region inside a continuous medium containing the points A and B. When the medium is deformed, A and B move to the positions A' and B'. The location of B relative to A prior to distortion is $\mathbf{D} = \mathbf{X}_B - \mathbf{X}_A$ and after distortion is $\mathbf{D}' = \mathbf{X}'_A - \mathbf{X}'_B$. The relative displacement of A and B is

$$\Delta\mathbf{D} = \mathbf{D}' - \mathbf{D}$$

$\mathbf{D}$, $\mathbf{D}'$, and $\Delta\mathbf{D}$ are all vectors. We characterize the gross distortion of the medium by the nine ratios of vector components

$$\begin{array}{lll} \dfrac{\Delta D_1}{D_1}, & \dfrac{\Delta D_1}{D_2}, & \dfrac{\Delta D_1}{D_3} \\ \dfrac{\Delta D_2}{D_1}, & \dfrac{\Delta D_2}{D_2}, & \dfrac{\Delta D_2}{D_3} \\ \dfrac{\Delta D_3}{D_1}, & \dfrac{\Delta D_3}{D_2}, & \dfrac{\Delta D_3}{D_3} \end{array}$$

To characterize the distortion at a point, we require the distance between A and B to become arbitrarily small, in which case the nine ratios approach the quantities

$$\begin{matrix} \dfrac{\partial D_1}{\partial X_1'} & \dfrac{\partial D_1}{\partial X_2'} & \dfrac{\partial D_1}{\partial X_3} \\ \dfrac{\partial D_2}{\partial X_1'} & \dfrac{\partial D_2}{\partial X_2'} & \dfrac{\partial D_2}{\partial X_3} \\ \dfrac{\partial D_3}{\partial X_1'} & \dfrac{\partial D_3}{\partial X_2'} & \dfrac{\partial D_3}{\partial X_3} \end{matrix}$$

as a limit. These are the nine components of the dyad $\nabla\mathbf{D}$, called the *distortion dyad;* we shall give it the symbol $\mathbf{M}$.

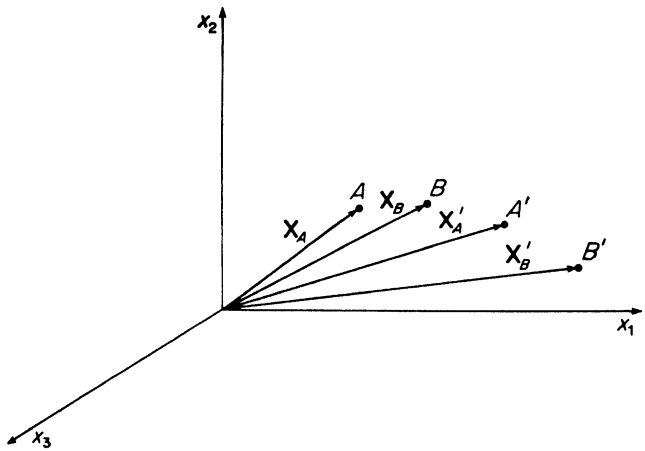


Figure 8.1 Locations of a pair of points in a medium before and after distortion.

Geometrically the diagonal terms of $\mathbf{M}$ represent an elongation of the medium per unit distance in the direction of elongation, while the cross terms represent a relative displacement of the points in the medium in a direction perpendicular to a component of the initial relative location of the points. The distortions of an infinitesimal portion of the medium corresponding to these are shown in Fig. 8.2.

Notice in Fig. 8.2 that if lines were originally drawn vertically on the rectangle, these would have been rotated by the distortion, whereas horizontal lines would not. However, if there had been an equal distortion, $\partial D_j/\partial X_i$, the sum of the angular displacements of all of the points in the rectangle would have been zero. When the distortion dyad is symmetric, it will represent a pure distortion, but otherwise it represents a distortion plus a rotation. We can, however, always break the dyad $\mathbf{M}$ into symmetric and antisymmetric

parts by the methods of p. 43. In this manner, we may write

$$\mathbf{M} = \boldsymbol{\varepsilon} + \mathbf{R}$$

where

$$\epsilon_{ij} = \frac{1}{2}\left\{\frac{\partial D_i}{\partial X_j} + \frac{\partial D_j}{\partial X_i}\right\} \tag{8.1}$$

and

$$R_{ij} = \frac{1}{2}\left\{\frac{\partial D_i}{\partial X_j} - \frac{\partial D_j}{\partial X_i}\right\} \tag{8.2}$$

R is an antisymmetric tensor representing an infinitesimal angular displacement; $\boldsymbol{\varepsilon}$ is a symmetric tensor called the *strain tensor*. $\boldsymbol{\varepsilon}$ will represent the

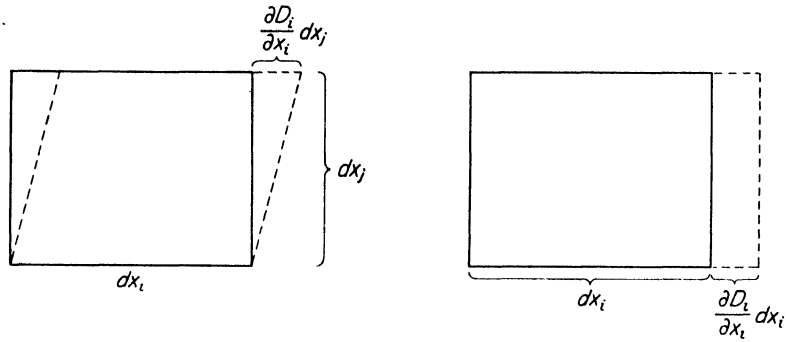


Figure 8.2 Schematic of strain of an element of volume in a continuous medium.

distortion of the medium just as adequately as **M**, but has the added advantage of always being symmetric. This is sufficiently convenient so that we will tolerate the one disadvantage in $\boldsymbol{\varepsilon}$. If the medium is distorted in such a way that $\mathbf{R} = 0$, then the shearing strains will be represented directly by the nondiagonal components of **M**, whereas the strain tensor $\boldsymbol{\varepsilon}$ is incapable of representing shearing strains directly because the nondiagonal components of $\boldsymbol{\varepsilon}$ are hybrid terms involving two different shear strain components. In the special case where

$$M_{ij} = M_{ji} \qquad (j \neq 1)$$

then M_{ij} represents a shear strain, and from Eq. (8.1)

$$\epsilon_{ij} = M_{ij}$$

8.2 Elasticity

Whenever a body is subjected to a stress, it will deform in a manner determined by the components of the stress tensor and the properties of the body. If a linear relationship exists between the stress and strain components, the material is said to be *elastic*, and the constants involved in the linear equations

are the elastic properties of the body. There are a variety of related methods of representing the elastic properties, each of which may be very useful in some applications and very cumbersome in others. The major objective of this section is to define these precisely and to show their relationship to one another.

The elastic compliance and elastic rigidity. Following along the line of the development of Chap. 6, the basic elastic properties are the elastic rigidity $\partial\mathfrak{d}/\partial\mathfrak{e}$ and the elastic compliance $\partial\mathfrak{e}/\partial\mathfrak{d}$. If the materials under consideration are not piezomagnetic or piezoelectric and have thermal expansion coefficients whose squares are negligible in comparison with the elastic compliance, then $\partial\mathfrak{d}/\partial\mathfrak{e}$ and $\partial\mathfrak{e}/\partial\mathfrak{d}$ become the inverse of one another. Otherwise, they must be related through a matrix equation such as that shown on p. 119. We shall consider only the simple case here.

Both $\partial\mathfrak{d}/\partial\mathfrak{e}$ and $\partial\mathfrak{e}/\partial\mathfrak{d}$ are rank four tensor properties (see the end of Chap. 3). Thus, $\partial\mathfrak{d}/\partial\mathfrak{e}$ and $\partial\mathfrak{e}/\partial\mathfrak{d}$ are not reciprocals in the sense that scalars are, even though they bear an inverse relationship to one another. It is actually very cumbersome to form the inverse of a rank four tensor, and we shall show the relationship between the components of $\partial\mathfrak{d}/\partial\mathfrak{e}$ and $\partial\mathfrak{e}/\partial\mathfrak{d}$ indirectly through two sets of elastic constants defined in the next section.

Denoting the rigidity tensor, $\partial\mathfrak{d}/\partial\mathfrak{e}$, by $\mathbf{C}$ and $\partial\mathfrak{e}/\partial\mathfrak{d}$ by $\mathbf{S}$, we may relate the stresses and strains by either of the set of equations

$$\epsilon_{ij} = S_{ijkl}\sigma_{kl} \tag{8.3}$$

$$\sigma_{ij} = C_{ijkl}\epsilon_{kl} \tag{8.4}$$

Both of these involve 81 equations, each of which contains 81 terms. However, both $\mathfrak{d}$ and $\mathfrak{e}$ are symmetric tensors. Consequently, from the first condition we may write

$$\epsilon_{ij} = S_{ijkl}\sigma_{kl} = S_{ijlk}\sigma_{lk}$$

and therefore

$$S_{ijkl} = S_{ijlk}$$

Similarly, the symmetry of $\mathfrak{e}$ requires that

$$S_{ijkl} = S_{jikl}$$

Therefore, only 36 of the components of the compliance tensor are distinct. The same is also applicable to the components of the rigidity tensor.

We may reduce the number of distinct components still further by considering the following argument. If U is the internal energy per unit volume, then

$$\sigma_{ij} = \frac{\partial U}{\partial \epsilon_{kl}}$$

but

$$S_{ijkl} = \frac{\partial \sigma_{ij}}{\partial \epsilon_{kl}} = \frac{\partial^2 U}{\partial \sigma_{ij}\, \partial \epsilon_{kl}}$$

However, since U is a perfect differential

$$S_{ijkl} = \frac{\partial^2 U}{\partial \sigma_{ij}\, \partial \epsilon_{kl}} = \frac{\partial^2 U}{\partial \epsilon_{kl}\, \partial \sigma_{ij}} = S_{klij}$$

The net effect of the three conditions

$$S_{ijkl} = S_{ijlk}, \qquad S_{ijkl} = S_{jikl}, \qquad S_{ijkl} = S_{klij} \tag{8.5a}$$

$$C_{ijkl} = C_{ijlk}, \qquad C_{ijkl} = C_{jikl}, \qquad C_{ijkl} = C_{klij} \tag{8.5b}$$

is to reduce the number of distinct components of the **S** and **C** tensors to 21. Even these are not totally independent, since by a proper choice of coordinates, all but 18 would become identically zero. However, in a coordinate system chosen by the convention of Sec. 7.5, there would be 21 distinct components. Of course, only the most asymmetric type of crystal will have this number of distinct components. Many become equal to each other or vanish for structures of symmetry higher than monoclinic.

The effect of crystal symmetry upon the compliance and rigidity components is given in Appendix B for the various crystal classes. However, we will illustrate how the information presented there was derived. We shall use crystal class 10 as an example. The symmetry operator defining this class is 4, where the x_3 axis is the axis of symmetry. The matrices for rotation through 180° and 90° respectively about the x_3 axis are

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The general formula for the transformation of a rank four tensor is

$$C'_{rstu} = a_{ri} a_{sj} a_{tk} a_{ul} C_{ijkl} \tag{8.6}$$

Whenever all of the nondiagonal terms of the transformation matrix are zero, 80 of the 81 terms on the righthand side of Eq. (8.6) will also be zero. The remaining term will have the same subscripts as that on the left. Thus, the first transformation will produce the result

$$C'_{rstu} = a_{r(r)} a_{s(s)} a_{t(t)} a_{u(u)} C_{rstu}$$

The product of the a's will be either $+1$ or -1. In the latter case, C'_{rstu} equals $-C_{rstu}$ equals 0 by Neumann's principle. This will occur for any combination of the a's in the following groups

$$a_{11}a_{11}a_{11}a_{33}, \qquad a_{22}a_{22}a_{22}a_{33}, \qquad a_{11}a_{11}a_{22}a_{33}$$
$$a_{11}a_{22}a_{22}a_{33}, \qquad a_{11}a_{33}a_{33}a_{33}, \qquad a_{22}a_{33}a_{33}a_{33}$$

From this and the conditions in Eq. (8.5b) we may deduce

$$C_{1113} = C_{1311} = C_{1131} = C_{3111} = 0$$
$$C_{2223} = C_{2322} = C_{2232} = C_{3222} = 0$$
$$C_{1123} = C_{2311} = C_{3211} = C_{1132} = 0$$
$$C_{1333} = C_{3133} = C_{3313} = C_{3331} = 0$$
$$C_{1223} = C_{2123} = C_{1232} = C_{2132} = C_{2312} = C_{2321} = C_{3212} = C_{3221} = 0$$
$$C_{2231} = C_{2213} = C_{1322} = C_{3122} = 0$$
$$C_{2333} = C_{3233} = C_{3323} = C_{3332} = 0$$
$$C_{2113} = C_{2131} = C_{1213} = C_{1231} = C_{1321} = C_{3121} = C_{1312} = C_{3112} = 0$$

We have now reduced the number of independent coefficients from 21 to 13 (and incidentally have deduced which of the tensor components are zero for crystal classes 3, 4, and 5). Of the remaining thirteen sets of equal components, we may select

$$\begin{array}{lllll} C_{1111}, & C_{1122}, & C_{1112}, & C_{1212}, & C_{1133} \\ C_{1313}, & C_{2222}, & C_{2212}, & C_{2233}, & C_{2313} \\ & C_{2323}, & C_{3312}, & C_{3333} & \end{array}$$

to test under the transformation

$$\begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

When $$C'_{1111} = a_{1i}a_{1j}a_{1k}a_{1l}C_{ijkl}$$

the only nonzero combination of a's occurs where the second subscript is 2 for all of the a's. Hence

$$C'_{1111} = C_{2222} = C_{1111}$$

Use of the appropriate transformation formulas produces

$$C'_{2222} = C_{1111} = C_{2222}$$

and $$C'_{3333} = C_{3333}$$

By a similar line of reasoning, we may deduce

$$\begin{array}{lll} C_{1112} = -C_{2221}, & C_{1133} = C_{2233}, & C_{2323} = C_{1313} \\ C_{3312} = 0, & C_{2313} = 0, & C_{2213} = C_{1122} \end{array}$$

Applying the conditions of Eq. (8.5b) to these, we obtain

$$C_{1112} = C_{1121} = C_{2111} = C_{1211} = -C_{2221} = -C_{2212} = -C_{1222} = -C_{2122}$$
$$C_{3312} = C_{3321} = C_{2133} = C_{1233} = 0$$

$$C_{2323} = C_{2332} = C_{3223} = C_{3232} = C_{1313} = C_{3113} = C_{1331} = C_{3131}$$
$$C_{2331} = C_{1323} = C_{2313} = C_{3213} = C_{3231} = C_{3123} = C_{1332} = C_{3132} = 0$$

We shall not prove it, but inversion after a rotation of 90° or 180° will not impose any further conditions upon the rigidity components. Thus, for this crystal class, of the 21 sets of distinct components, ten equal zero, three pairs are equal to one another, and one pair consists of one set equal to the negative of the other set. The remaining three are completely independent. A similar argument would apply to the components of the compliance tensor.

8.3 The elastic constants

Even the unperceptive reader must have noticed that 81 tensor components are rather cumbersome to handle. To avoid this, we might reason that six stress components and six strain components in an elastic medium could be interrelated by a set of six linear equations with a total of 36 constants. These should suffice to characterize the elastic properties of any material. This, in fact, is the common method of expressing the elastic properties of anisotropic materials. The only essential reasons for characterizing them first in terms of the components of the compliance or rigidity tensor were to show their general relationship to other equilibrium properties and to show how crystal symmetry affects them. Once this point is understood, we may proceed to define a more convenient set of constants and to show their relationship to the rigidity components.

So long as we are taking a big step in reducing the burden of different constants, we might just as well make our notation a bit more concise too. Let us make the notational changes

$$\begin{aligned}
&\epsilon_{11} = \epsilon_1, \quad \epsilon_{23} = \epsilon_{32} = \epsilon_4/2, \quad \sigma_{11} = \sigma_1, \quad \sigma_{32} = \sigma_{23} = \sigma_4 \\
&\epsilon_{22} = \epsilon_2, \quad \epsilon_{13} = \epsilon_{31} = \epsilon_5/2, \quad \sigma_{22} = \sigma_2, \quad \sigma_{13} = \sigma_{31} = \sigma_5 \\
&\epsilon_{33} = \epsilon_3, \quad \epsilon_{12} = \epsilon_{21} = \epsilon_6/2, \quad \sigma_{33} = \sigma_3, \quad \sigma_{21} = \sigma_{12} = \sigma_6
\end{aligned} \tag{8.7}$$

Then the elastic constants are defined by the set of equations

$$\begin{aligned}
\sigma_1 &= C_{11}\epsilon_1 + C_{12}\epsilon_2 + C_{13}\epsilon_3 + C_{14}\epsilon_4 + C_{15}\epsilon_5 + C_{16}\epsilon_6 \\
\sigma_2 &= C_{21}\epsilon_1 + C_{22}\epsilon_2 + C_{23}\epsilon_3 + C_{24}\epsilon_4 + C_{25}\epsilon_5 + C_{26}\epsilon_6 \\
\sigma_3 &= C_{31}\epsilon_1 + C_{32}\epsilon_2 + C_{33}\epsilon_3 + C_{34}\epsilon_4 + C_{35}\epsilon_5 + C_{36}\epsilon_6 \\
\sigma_4 &= C_{41}\epsilon_1 + C_{42}\epsilon_2 + C_{43}\epsilon_3 + C_{44}\epsilon_4 + C_{45}\epsilon_5 + C_{46}\epsilon_6 \\
\sigma_5 &= C_{51}\epsilon_1 + C_{52}\epsilon_2 + C_{53}\epsilon_3 + C_{54}\epsilon_4 + C_{55}\epsilon_5 + C_{56}\epsilon_6 \\
\sigma_6 &= C_{61}\epsilon_1 + C_{62}\epsilon_2 + C_{63}\epsilon_3 + C_{64}\epsilon_4 + C_{65}\epsilon_5 + C_{66}\epsilon_6
\end{aligned} \tag{8.8}$$

These equations are more concisely expressed in the form of the matrix equation

$$(\sigma_i) = (C_{ij})(\epsilon_j) \tag{8.9}$$

The same thermodynamic arguments which lead to the relationship C_{rstu} equals C_{turs} will also lead to C_{ij} equals C_{ji}.

The relationship between the elastic constants C_{ij} and the components of the rigidity tensor may be shown by expanding

$$\sigma_{ij} = C_{ijkl}\epsilon_{kl}$$

and (remembering the notational change) comparing the results with the corresponding equation of the set (8.8). Thus, for example,

$$\sigma_{12} = C_{1211}\epsilon_{11} + C_{1222}\epsilon_{22} + C_{1233}\epsilon_{33} + 2C_{1232}\epsilon_{32} + 2C_{1231}\epsilon_{31} + 2C_{1212}\epsilon_{12}$$

Comparison with the last of the set of Eqs. (8.8) results in

$$C_{1211} = C_{61}, \quad C_{1222} = C_{62}, \quad C_{1233} = C_{63}$$
$$C_{1232} = C_{64}, \quad C_{1231} = C_{65}, \quad C_{1212} = C_{66}$$

In this manner, we may build the set of relationships of Table 8.1.

Table 8.1 *The relationship between the elastic constants and the components of the rigidity tensor*

$C_{11} = C_{1111}$ $C_{12} = C_{1122} = C_{2211}$ $C_{13} = C_{1133} = C_{3311}$
$C_{14} = C_{1123} = C_{1132} = C_{2311} = C_{3211}$
$C_{15} = C_{1113} = C_{1131} = C_{1311} = C_{3111}$
$C_{16} = C_{1112} = C_{1121} = C_{2111} = C_{1211}$
$C_{22} = C_{2222}$ $C_{23} = C_{2233} = C_{3322}$ $C_{33} = C_{3333}$
$C_{24} = C_{2232} = C_{2223} = C_{2322} = C_{3222}$
$C_{25} = C_{2213} = C_{2231} = C_{1322} = C_{3122}$
$C_{26} = C_{2212} = C_{2221} = C_{1222} = C_{2122}$
$C_{34} = C_{3323} = C_{2333} = C_{3332} = C_{3233}$
$C_{35} = C_{3313} = C_{3331} = C_{1333} = C_{3133}$
$C_{36} = C_{3312} = C_{3321} = C_{2133} = C_{1233}$
$C_{44} = C_{2323} = C_{2332} = C_{3223} = C_{3232}$
$C_{45} = C_{2313} = C_{2331} = C_{3213} = C_{3231} = C_{1323} = C_{1332} = C_{3123} = C_{3132}$
$C_{46} = C_{2312} = C_{2321} = C_{3212} = C_{3221} = C_{1223} = C_{2123} = C_{1232} = C_{2132}$
$C_{55} = C_{1313} = C_{1331} = C_{3113} = C_{3131}$
$C_{56} = C_{1312} = C_{3112} = C_{1321} = C_{3121} = C_{1213} = C_{1231} = C_{2113} = C_{2131}$
$C_{66} = C_{1212} = C_{2112} = C_{1221} = C_{2121}$

To relate the inverse elastic constants to the components of the compliance tensor, substitute S_{ij} for C_{ij} and S_{rstu} for C_{rstu} in this table. The constants S_{14} to S_{36} inclusive are divided by 2 except S_{33}. Beginning with S_{44}, the inverse constants are divided by 4.

We may also define another set of elastic constants, S_{ij}, such that

$$\epsilon_i = S_{ij}\sigma_j \tag{8.10}$$

where $$(S_{ij}) = (C_{pq})^{-1*}$$

The elements of (S_{ij}) are related to the components of the compliance tensor. We may illustrate the relationship by expanding

$$\epsilon_{ij} = S_{ijkl}\sigma_{kl}$$

and comparing it with the appropriate member of the set of Eqs. (8.10). Thus, for example,

$$\epsilon_5 = S_{51}\sigma_1 + S_{52}\sigma_2 + S_{53}\sigma_3 + S_{54}\sigma_4 + S_{55}\sigma_5 + S_{56}\sigma_6$$

and

$$\epsilon_{31} = S_{3211}\sigma_{11} + S_{3222}\sigma_{22} + S_{3233}\sigma_{33} + 2S_{3232}\sigma_{32} + 2S_{3231}\sigma_{31} + 2S_{3221}\sigma_{21}$$

Consequently, since $\epsilon_5 = 2\epsilon_{31}$

$$S_{51} = 2S_{3211}, \qquad S_{52} = 2S_{3222}, \qquad S_{53} = 2S_{3233}$$
$$S_{54} = 4S_{3232}, \qquad S_{55} = 4S_{3231}, \qquad S_{56} = 4S_{3221}$$

Table 8.1 gives the complete set of relationships between the elastic constants, S_{ij}, and the components of the compliance tensor. The factors of 2 and 4 which have appeared here are due to the notational changes of Eq. (8.7). These were chosen in such a way that the (C_{pq}) and (S_{ij}) matrices would be symmetric.

Finally, since we have related the components of the rigidity to the elements of the (C_{pq}) matrix and the components of the compliance to the (S_{ij}) matrix, and since the matrices are the inverse of one another we have also established a relationship between the components of the rigidity and compliance tensors.

8.4 Technical elastic moduli

When material properties are defined along the lines developed in Chap. 6, they emerge as a unified, interrelated network of well-defined mathematical entities which bear a precise relationship to each other and to the structural symmetry of the solids to which they refer. Often, however, they are difficult, and occasionally impossible, to evaluate by direct experiment. Consequently, we usually find that older methods of handling properties are still being used. The technical elastic moduli are in this group.

* In order to compute any one of the elements of the matrix (S_{ij}) in terms of the elements of (C_{pq}), refer to the method outlined in Chap. 2, Sec. 2.

Technical elastic moduli are based upon three parameters, the *Young's modulus* Y, the *shear modulus* G, and *Poisson's ratio* μ. These are capable of characterizing the elastic properties of anisotropic materials having orthorhombic or higher symmetry. In the case of orthorhombic symmetry, we choose an axis parallel to the crystallographic axis. Then, three Young's moduli are distinguished and are defined by the equations

$$Y_1 = \frac{\partial \sigma_1}{\partial \epsilon_1}\Bigg)_{\sigma_2,\ldots,\sigma_6 = \text{constant}} \tag{8.11a}$$

$$Y_2 = \frac{\partial \sigma_2}{\partial \epsilon_2}\Bigg)_{\sigma_1,\sigma_3,\ldots,\sigma_6 = \text{constant}} \tag{8.11b}$$

$$Y_3 = \frac{\partial \sigma_3}{\partial \epsilon_3}\Bigg)_{\sigma_1,\sigma_2,\sigma_4,\sigma_5,\sigma_6 = \text{constant}} \tag{8.11c}$$

This should be contrasted with the elastic constants

$$C_{11} = \frac{\partial \sigma_1}{\partial \epsilon_1}\Bigg)_{\epsilon_2,\ldots,\epsilon_6 = \text{constant}}$$

Poisson's ratios are defined by the equation

$$\mu_{ij} = -\frac{\partial \epsilon_i}{\partial \epsilon_j} \tag{8.12}$$

with all of the stress components except σ_j held constant. Here the subscripts i and j range from 1 to 3, shearing strains being excluded from the definition. For orthorhombic crystals there are six Poisson's ratios. However, the same thermodynamic arguments that require C_{pq} to be equal to C_{qp} are also operative here, and result in the restriction that

$$\frac{Y_j}{\mu_{ij}} = \frac{Y_i}{\mu_{ji}}$$

For crystals of this symmetry, we also distinguish three shear moduli which are identical with certain elastic constants

$$G_1 = C_{44}, \qquad G_2 = C_{55}, \qquad G_3 = C_{66} \tag{8.13}$$

Thus we have twelve elastic moduli, but only nine of these are independent.

For tetragonal symmetry,

$$Y_1 = Y_2 \qquad \text{and} \qquad \mu_{12} = \mu_{21}$$
$$G_1 = G_2 \qquad\qquad \mu_{31} = \mu_{32}$$

which reduces the total number of independent constants to five. Materials having this symmetry are isotropic on planes parallel to the x_1x_2 plane and G_1, Y_1, and μ_{12} are related by

$$G_1 = \frac{Y_1}{2\mu_{12} + 1} \tag{8.14}$$

This relationship is derived in most undergraduate strength of materials texts and need not be repeated here. Generally it is only valid for isotropic materials and can only be applied here because G_1, Y_1, and μ_{12} are properties measured on the transverse plane.

Materials having cubic symmetry have only three independent coefficients

$$Y = Y_1 = Y_2 = Y_3$$

$$\mu = \mu_{12} = \mu_{21} = \mu_{31} = \mu_{13} = \mu_{23} = \mu_{32}$$

$$G = G_1 = G_2 = G_3$$

Isotropic materials have only two, either Y and μ or Y and G, since these are related through

$$G = \frac{Y}{2\mu + 1}$$

The Y's and μ's are not tensor components, and the computation of a distortion from the technical elastic moduli and an arbitrary stress involves the specialized geometrical techniques peculiar to that area of mechanics known as mechanics of materials. In contrast, if the components of the rigidity tensor are known, any strain due to an arbitrary stress may be computed directly from a double contraction of the rigidity tensor and stress tensor. The treatment of piezoelectric or piezomagnetic materials is very awkward if one attempts it in terms of the technical elastic moduli. Despite these disadvantages, the technical elastic moduli have one outstanding virtue: they are readily amenable to direct measurement. Perhaps the significance of this will be clearer if the reader constructs a mental experiment to determine

$$C_{11} = \frac{\partial \sigma_1}{\partial \epsilon_1}$$

in which all the strains other than ϵ_1 are held constant, as contrasted with

$$Y_1 = \frac{\partial \sigma_1}{\partial \epsilon_1}$$

in which all the stresses other than σ_1 are held constant. It would be very difficult to arrange the boundary conditions to meet the requirements of the first equation, whereas this difficulty evaporates for the second case.

The elastic constants and the technical elastic moduli can be related through a procedure illustrated by the following example. In a crystal having orthorhombic symmetry

$$\sigma_1 = C_{11}\epsilon_1 + C_{12}\epsilon_2 + C_{13}\epsilon_3$$

$$\sigma_2 = C_{21}\epsilon_1 + C_{22}\epsilon_2 + C_{23}\epsilon_3$$

$$\sigma_3 = C_{31}\epsilon_1 + C_{32}\epsilon_2 + C_{33}\epsilon_3$$

since, according to Appendix B, C_{14}, C_{15}, C_{25}, C_{16}, C_{26}, and $C_{36} = 0$. If we differentiate the first of these with respect to ϵ_1, holding σ_2 and σ_3 constant, we obtain

$$Y_1 = \frac{\partial \sigma_1}{\partial \epsilon_1}\Bigg)_{\sigma_2,\sigma_3 \text{ constant}} = C_{11} - \mu_{21}C_{12} - \mu_{31}C_{13}$$

If we now differentiate the same equation with respect to ϵ_2, holding σ_1 and σ_3 constant, we obtain

$$0 = -\mu_{12}C_{11} + C_{12} - \mu_{32}C_{13}$$

A similar differentiation with respect to ϵ_3 produces

$$0 = -\mu_{13}C_{11} - \mu_{23}C_{12} + C_{13}$$

All of the technical elastic moduli may be obtained from direct measurement, and the elastic constants C_{11}, C_{12}, and C_{13} may be evaluated from a simultaneous solution of

$$\begin{aligned} Y_1 &= C_{11} - \mu_{21}C_{12} - \mu_{31}C_{13} \\ 0 &= -\mu_{12}C_{11} + C_{12} - \mu_{32}C_{13} \\ 0 &= -\mu_{13}C_{11} - \mu_{23}C_{13} + C_{13} \end{aligned}$$

In a similar manner, the other elastic constants may be related to the technical elastic moduli.

It is very easy to lose perspective in any detailed treatment of a single topic in which three different sets of quantities are used to characterize one mode of behavior. Let us now look backward to see why we have these different sets of quantities and where each will be most useful.

The technical elastic moduli are simply a set of constants which are defined in such a way that they are easy to measure experimentally.

The elastic compliance and rigidity tensors are not amenable to direct measurement, but they are well-defined mathematical entities which lend themselves to algebraic manipulation; given either of these and the applied stress, the strain can be computed by a simple contraction. Also, because the transformation of a tensor from one coordinate system to another is algebraically simple, we can best show the restrictions imposed by crystal symmetry upon the elastic properties by expressing the elastic properties in tensor form.

The elastic constants allow us to relate stress and strain by matrix operations in a less cumbersome way than by using the corresponding tensor quantities because we only have to work with 36 equations instead of 81. Each of these constants—the technical elastic moduli, the elasticity tensors, and the elastic constant matrices—has its own peculiar area of utility and convenience, but the latter two must be related to the first if we are to evaluate them experimentally. We have shown these relationships.

8.5 Dynamic strain and the speed of sound

The velocity of propagation of a distortion wave through a solid bears an important relationship to the elastic and thermal properties of materials. The relationship to thermal properties will be discussed in the next chapter. The primary emphasis of this section will be upon the relationship between the speed of sound in a solid and the elastic constants of the solid. Indeed, the foremost method of evaluating the elastic constants of anisotropic crystals is to measure the resonant and antiresonant frequencies of certain vibrational modes. These, of course, are simply standing distortion waves, and are identical with sound waves.

When a dynamic stress is applied to a surface of a solid, the distortion it produces will travel through the solid in such a complex way that in the general case, where an arbitrary stress is applied to an anisotropic body of arbitrary shape, it would probably defy analysis. In determining the elastic constants of anisotropic solids, the size, shape, and orientation of the crystallographic axes are carefully selected to simplify the combinations of possible strain waves and, consequently, to simplify the analysis.

The general partial differential equations that govern the propagation of distortion waves in a solid can be deduced in this way. Select a small region inside the solid not adjacent to the source of the disturbance. The forces acting upon the surface of the region will be elastic forces due to the local distortion of the medium. These will be balanced by the inertial forces of the mass points inside the region. This statement is formulated in the equation

$$\int_\tau \rho \ddot{\mathbf{D}}\, d\tau = \int_\Sigma (\mathbf{C} : \boldsymbol{\varepsilon}) \cdot d\mathbf{A}$$

By the divergence theorem (p. 17)

$$\int_\Sigma (\mathbf{C} : \boldsymbol{\varepsilon}) \cdot d\mathbf{A} = \int_\tau \nabla \cdot (\mathbf{C} : \boldsymbol{\varepsilon})\, d\tau$$

If $\Sigma \longrightarrow 0$ as $\tau \longrightarrow 0$, and if this relationship is valid for arbitrarily small τ, we may write

$$\rho \ddot{\mathbf{D}} = \nabla \cdot (\mathbf{C} : \boldsymbol{\varepsilon}) \tag{8.15}$$

If the density changes by only a negligible amount under a distortion

$$\nabla \rho \mathbf{D} = \rho \nabla \mathbf{D} = \rho \mathbf{M}$$

where $\mathbf{M}$ is the distortion dyad defined in Sec. 8.1. Thus, we may operate upon both sides of Eq. (8.15) with ∇ to produce

$$\rho \nabla \ddot{\mathbf{D}} = \nabla\nabla \cdot (\mathbf{C} : \boldsymbol{\varepsilon}) = \rho \ddot{\mathbf{M}} \tag{8.16}$$

Expanding the tensor equation into its components produces the following nine simultaneous partial differential equations governing the distortion of

the medium as a function of time and the coordinates

$$\rho \ddot{M}_{ji} = \frac{\partial^2}{\partial x_j \, \partial x_k} C_{kimn} \epsilon_{mn} \tag{8.17}$$

The solution of these depends upon the boundary conditions and it is here that, in determining the elastic constants, the size, shape, orientation of crystallographic axes, and mode of excitation are selected to simplify and reduce Eq. (8.17).

Since $\epsilon_{mn} = \epsilon_{nm}$, the summation on the righthand side of (8.17) will always contain three terms of the form

$$\frac{\partial^2}{\partial x_j \, \partial x_k} [2C_{kimn}\epsilon_{mn}]$$

and since

$$2\epsilon_{mn} = M_{mn} + M_{nm}$$

Equation (8.17) may be expressed entirely in terms of the components of the strain dyad, **M**. Also, since the diagonal components are identical for **M** and

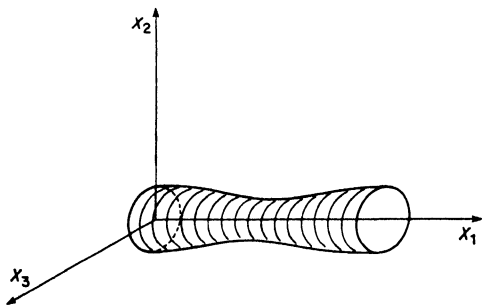


Figure 8.3 Compression wave in a rod.

ε, three of the equations may be expressed entirely in terms of the strain tensor components, ϵ_{ij}.

As an application of (8.17), we shall show the relationship between the isentropic elastic constants, C_{11} and C_{12}, of an isotropic medium and the velocity of sound. First we consider the propagation of a tension-compression wave in a longitudinal direction along a rod. Let us choose the x_1 axis along the axis of the rod and assume that the primary disturbance is an ϵ_{11} strain. Now if a material is stretched or compressed, its lateral dimensions will also change. (See Fig. 8.3.) Consequently, the primary strain wave will be accompanied by secondary ϵ_{22} and ϵ_{33} waves which move with the same velocity. There will also be a shear-distortion wave associated with the primary ϵ_{11} wave. Since ϵ_{11} is a function of x_1, ϵ_{22} and ϵ_{33} will also be. The entire combinations of dynamic strains is called a *longitudinal wave*. All of these come automatically out of Eq. (8.17). For the geometry we have chosen, we

need consider only the equation in which

$$M_{ij} = (\epsilon_{ij}) = M_{11} = \epsilon_{11}$$

which is, in extenso,

$$\begin{aligned}\rho \frac{\partial^2 \epsilon_{11}}{\partial t^2} = \frac{\partial^2}{\partial x_1^2} (C_{1111}\epsilon_{11} + C_{1122}\epsilon_{22} + C_{1133}\epsilon_{33} \\ + 2C_{1121}\epsilon_{21} + 2C_{1131}\epsilon_{31} + 2C_{1132}\epsilon_{32}) + \frac{\partial^2}{\partial x_1 \partial x_2} (C_{2111}\epsilon_{11} \\ + C_{2122}\epsilon_{22} + C_{2122}\epsilon_{22} + 2C_{2121}\epsilon_{21} + 2C_{2131}\epsilon_{31} + 2C_{2132}\epsilon_{32}) \\ + \frac{\partial^2}{\partial x_1 \partial x_3} (C_{3111}\epsilon_{11} + C_{3122}\epsilon_{22} + C_{3133}\epsilon_{33} + 2C_{3121}\epsilon_{21} \\ + 2C_{3131}\epsilon_{31} + 2C_{3132}\epsilon_{32}) \qquad (8.18)\end{aligned}$$

Consulting Appendix B, we see that for an isotropic medium many of the tensor components C_{ijkl} are zero and others equal. In terms of the elastic constants, (8.18) reduces to

$$\begin{aligned}\rho \frac{\partial^2 \epsilon_{11}}{\partial t^2} = \frac{\partial^2}{\partial x_1^2} [C_{11}\epsilon_{11} + C_{12}(\epsilon_{22} + \epsilon_{33})] \\ + \frac{\partial^2}{\partial x_1 \partial x_2} (C_{11} - C_{12})\epsilon_{21} + \frac{\partial^2}{\partial x_1 \partial x_3} (C_{11} - C_{12})\epsilon_{31} \qquad (8.19)\end{aligned}$$

Notice that in Fig. 8.3, the shear strain increases linearly in the x_2 and x_3 directions. Thus

$$\frac{\partial}{\partial x_2} (C_{11} - C_{12})\epsilon_{21}$$

is dependent only upon x_1. The same is true of

$$\frac{\partial}{\partial x_3} (C_{11} - C_{12})\epsilon_{31}$$

The rate of change of these with respect to x_1 will be negligible if the lateral dimensions of the rod are small in comparison with the wavelength of the disturbance. Under these conditions, Eq. (8.19) reduces to

$$\rho \frac{\partial^2 \epsilon_{11}}{\partial t^2} = \frac{\partial^2}{\partial x_1^2} [C_{11}\epsilon_{11} + C_{12}(\epsilon_{22} + \epsilon_{33})] \qquad (8.20)$$

The term inside the square brackets is equal to the technical elastic modulus Y times ϵ_{11}.

When the strain wave is periodic, the solution of Eq. (8.20) leads to a propagation velocity

$$v_1 = \sqrt{\frac{Y}{\rho}} \qquad (8.21)$$

Figure 8.4 is an exaggerated schematic diagram of the type of distortion that occurs with the passage of a transverse wave through a thin rod. The primary distortion is an M_{21} distortion accompanied by a small longitudinal contraction. In this case, again only one of the set of Eqs. (8.17) is necessary

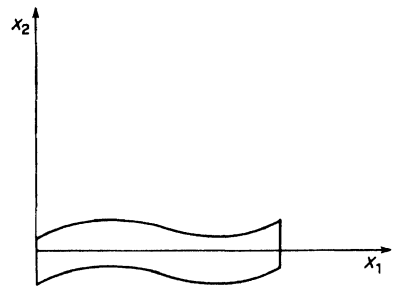


Figure 8.4 Shear wave in a rod.

to describe the disturbance. After expansion, and making the appropriate substitution from Appendix B for an isotropic medium, this equation reduces to

$$\rho \frac{\partial^2 M_{21}}{\partial t^2} = \frac{\partial^2}{\partial x_2 \, \partial x_1} [C_{11}\epsilon_{11} + C_{12}(\epsilon_{22} + \epsilon_{33})] + \frac{\partial^2}{\partial x_2^2} \left\{ \frac{(C_{11} - C_{12})}{2} (M_{12} - M_{21}) \right\} + \frac{\partial^2}{\partial x_2 \, \partial x_3} \left\{ \frac{C_{11} - C_{12}}{2} (M_{31} + M_{13}) \right\} \tag{8.22}$$

The tensile strains and their derivatives are negligible and the M_{21}, M_{31}, and M_{13} modes of deformation may be suppressed by the initial excitation of the wave. Thus (8.22) reduces to its final form

$$\rho \frac{\partial^2 M_{21}}{\partial t^2} = \frac{C_{11} - C_{12}}{2} \frac{\partial^2 M_{12}}{\partial x_2^2} \tag{8.23}$$

Just as in Eq. (8.19), the solution of this equation leads to a propagation velocity

$$v_t = \sqrt{\frac{C_{11} - C_{12}}{2}} = \sqrt{\frac{G}{\rho}}$$

By measuring the velocity of longitudinal and transverse waves we evaluate Y and G. From these, μ can be computed and finally, using the methods developed in Sec. 8.4, we may evaluate C_{11} and C_{12}.

Table 8.2 lists the room temperature elastic constants of a number of single crystal materials. It also shows the variation of the elastic constant when one material is alloyed with an isomorph. These values were obtained from acoustic measurements of single crystals.

Table 8.2 *Room temperature elastic constants of various materials**

Dyne $cm^{-2} \times 10^{-12}$

Material	Structure	C_{11}	C_{22}	C_{33}	C_{44}	C_{55}	C_{66}	C_{12}	C_{13}	C_{23}	C_{14}
U	Orthorhombic	2.15	1.99	2.67	1.24	0.734	0.743	0.465	0.218	1.076	
Mg	Hcp	0.585		0.610	0.166			0.250	0.208		
Bi	Trigonal	1.88		0.44	1.08			−0.90	0.241		0.42
Cu	fcc	1.679			0.751			1.213			
Ni	fcc	2.500			1.235			1.500			
S	Orthorhombic	0.024	0.0205	0.0483	0.043	0.087	0.076	0.133	0.177	0.159	
Cds		0.843		0.939	0.149		0.161	0.521	0.464		
GaAs	Cubic	1.188						0.538	0.594		
AlSb	Cubic	0.894						0.442	0.415		
GaSb	Cubic	0.885						0.404	0.432		
InSb	Cubic	0.672						0.367	0.302		
Topaz	Orthorhombic	2.82	3.49	2.95	1.08	1.33	1.31	1.26	0.85	0.88	
$CaCO_3$	Orthorhombic	1.60	0.87	0.85	0.41	0.25	0.43	0.37	0.17	0.15	
$BaSO_4$	Orthorhombic	0.862	0.917	0.108	0.126	0.287	0.244	0.523	0.34	0.356	
Rochelle salt	Orthorhombic	0.362	0.472	0.547	0.127	0.031	0.098	0.231	0.273	0.269	

* Compiled from the following sources: Bolef and Menes, *J. Applied Phys.*, **38**, 1410 (1960); S. K. Joshi, *Phys. Rev.*, **121**, 41 (1961); Smenk and Smith, *Acta Metal.*, **8**, 398 (1960); Fisher and McSkimm, *J. Applied Phys.*, 29, (1958).

8.6 Plasticity

In general a material may respond to a stress in various ways, and the response that is described by the term *elastic* will occur only in certain ranges of stress, if it occurs at all. Elasticity refers to cases where a functional relationship exists between the stress tensor and the strain tensor. Such a functional relationship occurs only for the rather specialized conditions discussed in Chap. 6 that allow us to characterize the properties of a material as compliances or rigidities. It will be recalled that in such cases all properties are defined in terms of reversible infinitesimal changes in a pair of parameters between two equilibrium states. In dealing specifically with the deformation resulting from a stress, there are two rather broad categories of commonly occurring behavior that do not conform to these restrictions. One of these processes is called *creep*, and refers to cases where the strain varies with time while the stress level is held constant. The other category refers to situations in which the strain does not return to zero when the stress is reduced to zero. Deformations of this type are called *plastic*. Most materials behave elastically, or almost elastically, at low stress levels. However, if an applied stress is increased montonically, eventually a stress level is attained at which the process becomes irreversible, and plastic deformation begins to occur. The stress at which this occurs is called the *elastic limit* of the material.

The elastic properties of a material are said to be structurally insensitive. This means that the elastic constants are relatively insensitive to factors such as grain size, previous mechanical or thermal history, small changes in chemical composition, and the like. However, the plastic behavior of a material will be quite sensitive to these factors, and chemically identical materials may be strikingly different in their plastic behavior. In particular the aforementioned factors will have a very strong effect upon the value of the elastic limit.

Our objective here is to consider the mechanisms by which plastic deformation occurs, and to correlate this mechanism with the observed effects of thermal and mechanical history, grain size, and like external factors upon the elastic limit. As the topic is developed we shall find that many different factors such as grain boundaries, dislocation density, and orientation of crystallographic planes relative to stress components are involved.

8.7 Modes of irreversible deformation; slip and twinning

In a geometrical sense, there are only three possible ways that a single crystal can be permanently deformed by the action of a stress: separation of the atoms across a plane of maximum tensile stress, lateral displacement of the atoms on each side of a crystallographic plane with respect to one another,

and twinning. The first of these is never observed and there are good theoretical reasons for its absence.

The second process, called *slip*, occurs when a shear stress acts upon crystallographic planes causing all of the atoms on one side of the plane to be

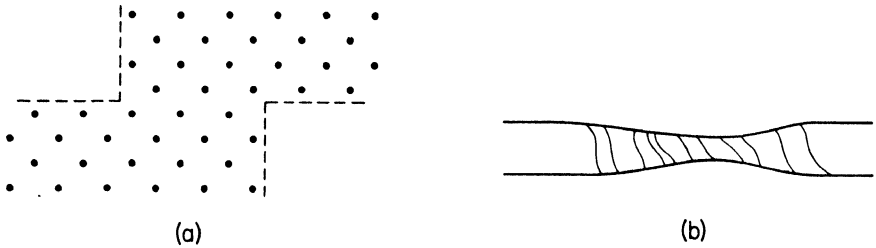


Figure 8.5 (a) Arrangement of atoms on each side of a slip plane. (b) Sketch of glide lines in a single crystal.

displaced laterally with respect to those on the other side of the plane. After displacement, the atoms of the crystal are still in registry with one another as shown in Fig. 8.5a. When the process has become extensive, the occurrence of slip may be detected visually from the development of glide lines on the surface of the crystal. Figure 8.5b shows a crystal in which these glide lines have developed under an axial load.

Whereas slip is due primarily to the action of a homogeneous shear, the third mode of deformation, *twinning*, occurs under an inhomogeneous shear.

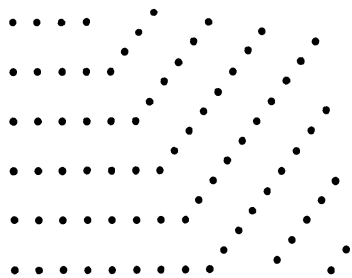


Figure 8.6 Schematic of the arrangement of the atoms on each side of a twin boundary.

Figure 8.6 shows the internal arrangement of the atoms in a twinned crystal and Figure 8.7 shows the external appearance of such a crystal.

Although both slip and twinning may occur in a polycrystalline material when it is stressed beyond its elastic limit, the plastic behavior of such a material is governed primarily by the slip process. The reason is that slip

occurs at a lower stress level than does twinning. Twins are formed principally at sites where slip in one crystal is inhibited by neighboring crystals oriented in such a way that they are more capable of withstanding the stress.

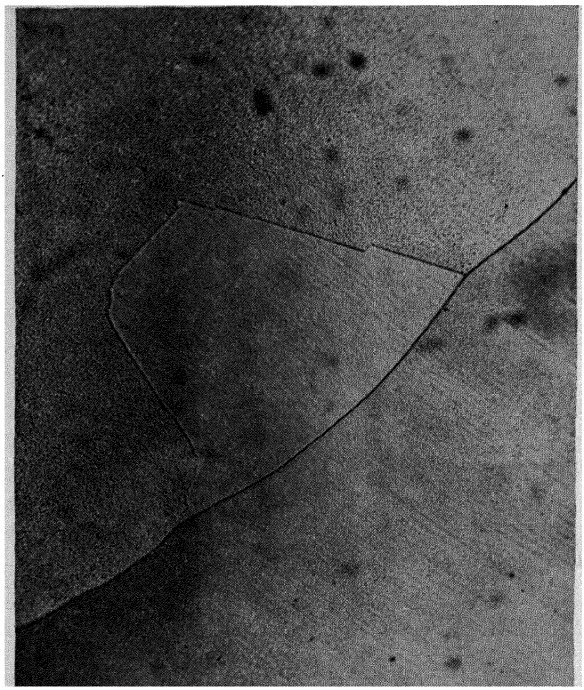
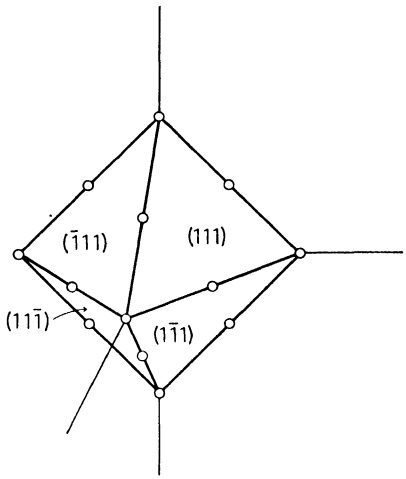


Figure 8.7 Annealing twin in Fe-Al alloy. (After Cahn & Coll, *Acta Metallurgica* **9**, 138–148 (1961). Photograph courtesy of R. W. Cahn, Department of Physical Metallurgy, University of Birmingham.)

8.8 Slip systems in single crystals and the critical shear stress

When slip is observed in a single crystal, it is found invariably that the slip direction coincides with one of several equivalent directions having the highest linear density of atoms in the crystal. Also it is almost invariably true that the slip plane will be one of the crystallographic planes having the highest areal density of atoms. Ignoring the "almost" in the previous sentence for the present, we can look at the basic cell of any crystal class and pick out the combinations of planes and directions in which slip can occur. This set of combinations is called a *slip system*. For example, in the fcc class there are twelve such combinations meeting the requirement of highest areal density on the plane and highest linear density of atoms. In Table 8.3 the planes of highest areal density are tabulated with their respective directions of highest linear density for fcc crystals.

Table 8.3 *Slip system of an fcc crystal*

External appearance	Direction of highest linear density	Planes of highest areal density	
	$\langle 1\bar{1}0 \rangle$	(111)	$(11\bar{1})$
	$\langle 10\bar{1} \rangle$	(111)	$(1\bar{1}1)$
	$\langle 01\bar{1} \rangle$	(111)	$(\bar{1}11)$
	$\langle 101 \rangle$	$(\bar{1}11)$	$(11\bar{1})$
	$\langle 110 \rangle$	$(\bar{1}11)$	$(1\bar{1}1)$
	$\langle 011 \rangle$	$(11\bar{1})$	$(1\bar{1}1)$

Let us imagine that all of the components of a stress are uniformly increased from zero until the onset of slip occurs. At this point, the value of the stress component resolved on the slip plane in the slip direction is called the *critical shearing stress*. If the components of the applied stress are known at the onset of slip, and if the slip plane and slip direction are known, the critical shear stress can be computed. Let $\mathbf{\sigma}_s$ represent the stress level at the onset of slip and let $\mathbf{n}$ be a unit vector normal to the plane of slip. Then if A is the area of the slip plane, the total force acting upon the slip plane will be

$$\mathbf{F} = \mathbf{\sigma}_s \cdot A\mathbf{n}$$

If we let $\mathbf{v}$ be a unit vector in the slip direction, then the component of the force tending to produce slip will be

$$\mathbf{v} \cdot \mathbf{\sigma}_s \cdot A\mathbf{n}$$

and the resolved shear stress acting upon this plane in the slip direction will be

$$\sigma_r = \frac{\mathbf{v} \cdot \mathbf{\sigma}_s \cdot A\mathbf{n}}{A} = \mathbf{v} \cdot \mathbf{\sigma}_s \cdot \mathbf{n}$$

Example 8.1. The longitudinal axis of a rectangularly shaped single crystal lies in the $\langle 111 \rangle$ crystallographic direction. The basic cell of this crystal is fcc. The crystal is subjected to a tensile load on the longitudinal axis which is increased gradually to 1000 psi, at which point slip is observed. Find the crystallographic plane or planes upon which slip would have most likely occurred, and compute the critical shearing stress for the crystal.

In a coordinate system with the x_1 axis aligned with the longitudinal axis of the crystal, the only nonzero stress component at the onset of slip is $\sigma_{11} = 1000$.

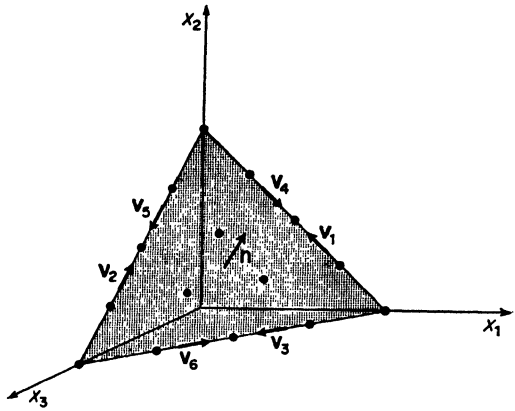


Figure 8.8 One of a set of planes in an fcc crystal having highest areal density of atoms. The unit vector n is normal to this plane, while the unit vectors designated with subscript v indicate possible slip directions. These correspond to directions of highest linear density of atoms. In the coordinate system indicated above, the various unit vectors have components tabulated below:

	x_1	x_2	x_3
$\mathbf{n}$	$\frac{\sqrt{3}}{3}$	$\frac{\sqrt{3}}{3}$	$\frac{\sqrt{3}}{3}$
$\mathbf{v}_1$	$-\frac{\sqrt{2}}{2}$	$\frac{\sqrt{2}}{2}$	0
$\mathbf{v}_2$	0	$\frac{2}{2}$	$-\frac{\sqrt{2}}{2}$
$\mathbf{v}_3$	$-\frac{\sqrt{2}}{2}$	0	$\frac{\sqrt{2}}{2}$
$\mathbf{v}_4$	$\frac{\sqrt{2}}{2}$	$-\frac{\sqrt{2}}{2}$	0
$\mathbf{v}_5$	0	$-\frac{\sqrt{2}}{2}$	$\frac{\sqrt{2}}{2}$
$\mathbf{v}_6$	$\frac{\sqrt{2}}{2}$	0	$-\frac{\sqrt{2}}{2}$

Transforming to the crystallographic axes, the stress components are

$$\sigma'_{ij} = a_{ik}a_{jl}\sigma_{kl}$$

$$\sigma'_{11} = \frac{1000}{3}, \quad \sigma'_{12} = \frac{1000}{3}, \quad \sigma'_{13} = \frac{1000}{3}$$

$$\sigma'_{22} = \frac{1000}{3}, \quad \sigma'_{23} = \frac{1000}{3}, \quad \sigma'_{33} = \frac{1000}{3}$$

The resolved shear stress for each member of the slip system is

$$\sigma_r = \mathbf{v} \cdot \mathbf{\sigma} \cdot \mathbf{n} = v_i \sigma_{ij} n_j$$

For the $\langle 1\bar{1}0 \rangle$ direction, $v_1 = \sqrt{2}/2$, $v_2 = -\sqrt{2}/2$, and $v_3 = 0$. For the $(11\bar{1})$ plane, $n_1 = \sqrt{3}/3$, $n_2 = \sqrt{3}/3$, and $n_3 = -\sqrt{3}/3$. Hence,

$$\sigma_r = \frac{1000}{3}\left(\frac{\sqrt{6}}{6} + \frac{\sqrt{6}}{6} - \frac{\sqrt{6}}{6} - \frac{\sqrt{6}}{6} + \frac{\sqrt{6}}{6}\right) = 0$$

In a similar manner, we compute

$\langle 101 \rangle$ $(11\bar{1})$	$\sigma_r = 1000\sqrt{6}/18$
$\langle 011 \rangle$ $(11\bar{1})$	$\sigma_r = 1000\sqrt{6}/18$
$\langle 1\bar{1}0 \rangle$ (111)	$\sigma_r = 0$
$\langle 10\bar{1} \rangle$ (111)	
$\langle 01\bar{1} \rangle$ (111)	
$\langle 10\bar{1} \rangle$ $(1\bar{1}1)$	$\sigma_r = 0$
$\langle 110 \rangle$ $(1\bar{1}1)$	$\sigma_r = 1000\sqrt{6}/18$
$\langle 011 \rangle$ $(1\bar{1}1)$	$\sigma_r = 1000\sqrt{6}/18$
$\langle 01\bar{1} \rangle$ $(\bar{1}11)$	$\sigma_r = 0$
$\langle 110 \rangle$ $(\bar{1}11)$	$\sigma_r = 1000\sqrt{6}/18$
101 $(\bar{1}11)$	$\sigma_r = 1000\sqrt{6}/18$

With the load applied in the manner specified, six of the members of the slip system are subjected to zero stress, and the others sustain equal stresses. One of these suffers slip when the resolved shear reaches $1000\sqrt{6}/18$ psi. Just which one will slip could be determined by any slight misalignment of the load, which would shift a part of the stress off some of these and increase it on others. When slip does begin, a further misalignment will occur successively upon all of the planes which sustain approximately equal stresses.

The reader will recall our saying earlier that almost invariably the slip system consists of planes of highest areal density and directions of highest linear density. Occasionally slip systems also develop on planes having the second highest areal density as well. In general, the critical shearing stress for this second slip system will be higher than that of the first. However, if the ratio of the critical shear stress of the second slip system to that of the first slip system exceeds a certain value, then it will be impossible for slip ever to develop on the second slip system. By way of example, we compute this critical ratio for the fcc crystal.

The second slip system in the fcc crystal consists of the following planes and directions:

$\langle 0\bar{1}1 \rangle$	$\langle 101 \rangle$	$\langle 110 \rangle$
(100)	(010)	(001)
$\langle 01\bar{1} \rangle$	$\langle 10\bar{1} \rangle$	$\langle \bar{1}10 \rangle$

Let us choose a stress condition to produce maximum resolved shear on the $(001)\langle\bar{1}10\rangle$ member of the second slip system, and compute the stress that this imposes upon the most unfavorably oriented member of the first slip system.

Choosing a coordinate system x_i coincident with the crystallographic axes, a pure tension T in the x_1 direction will produce a maximum resolved shear of magnitude $T/2$ on the (001) plane in the $\langle 1\bar{1}0\rangle$ direction. In order to compute the shear component of this stress on the (111) plane in the $\langle 1\bar{1}0\rangle$, we form the double contraction

$$\sigma_r = \mathbf{u}\cdot\boldsymbol{\sigma}\cdot\mathbf{v}$$

where $\mathbf{v}$ is a unit vector in the $\langle 1\bar{1}0\rangle$ direction and $\mathbf{u}$ is a unit vector perpendicular to the (111) plane. The components of $\mathbf{u}$, $\boldsymbol{\sigma}$, and $\mathbf{v}$ are

$$u_1 = u_2 = u_3 = \frac{\sqrt{3}}{3}, \qquad \sigma_{11} = T$$

$$v_1 = \frac{\sqrt{2}}{2}, \qquad v_2 = -\frac{\sqrt{2}}{2}, \qquad v_3 = 0$$

All other components of $\boldsymbol{\sigma}$ are zero. Hence

$$\sigma_r = \frac{\sqrt{6}\,T}{6}$$

Here we have deliberately arranged our stress conditions to attempt to produce slip on the (001) planes, but we are not able to avoid producing stress components on the (111) planes. The ratio of these is

$$\frac{T/2}{\sqrt{6}\,T/6} = \frac{3}{\sqrt{6}} = \frac{\sqrt{6}}{2}$$

Now if the ratio of the critical shearing stress for the (001) planes exceeds that of the (111) by a factor greater than or equal to $\sqrt{6}/2$, then there will be no possibility of the second slip system's becoming operative. On the other hand, if the critical shear stress for the second slip system is less than $\sqrt{6}/2$ times as great as the critical shear stress for the first slip system, both systems may be operative depending upon the orientation of the system which is subjected to a critical resolved shear under an arbitrary applied stress.

8.9 Critical shear stress of perfect single crystals

Historically, metallurgists and solid state physicists had long been puzzled over the low value of critical shear stress exhibited by single crystals. The energy of the bonds between the atoms of a crystal can be obtained directly from the heat of sublimation, and the shape of the potential energy versus distance curve can be computed from quantum mechanics. Assuming that

in the slip process all of the atoms on one side of a slip plane move simultaneously, the critical shear stress can be computed from the summation of the bonding forces between the atoms on each side of the slip plane when the relative displacement is half the interatomic distance in the slip direction.

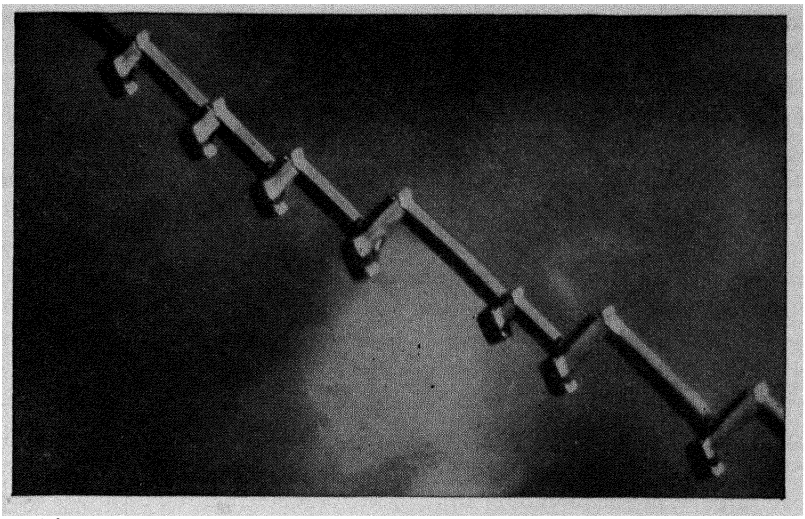
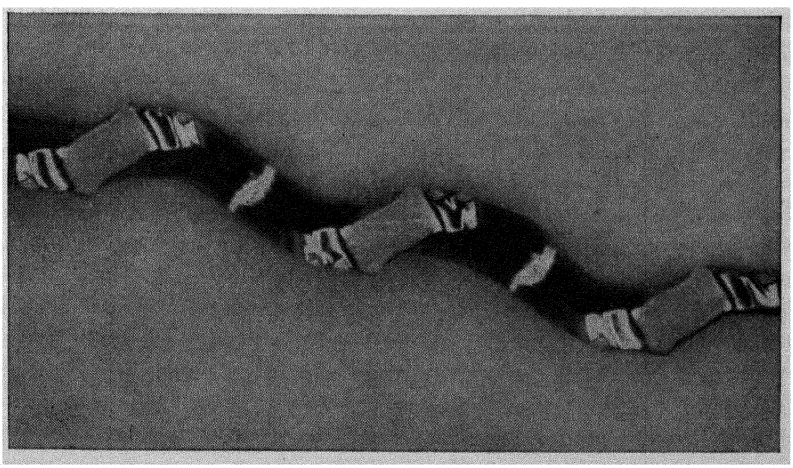


Figure 8.9 Photomicrographs of iron whiskers. (Courtesy of Dr. R. W. De Bois, General Electric Research Laboratory, Schenectady, N. Y.)



Such a computation gives values of critical shear stress that are on the order of 1000 times higher than those that are actually observed in single crystals.

In order to account for this discrepancy, the presence of dislocations in the crystal was postulated. In its original development, there was no direct evidence of the existence of dislocations. However, dislocation theory is capable

of providing a qualitative explanation of many aspects of plastic behavior, and it was widely accepted despite the lack of direct evidence. We shall show how it is applied to various aspects of plasticity in the next few sections.

Since the introduction of dislocation theory, there have been two major developments in support of it. Using electron microscope techniques, it has been possible to photograph the growth rings associated with spiral dislocations. The other development is the discovery of metal whiskers.

If protected from external mechanical disturbances, many metals will grow whiskers spontaneously. These are tiny single crystals that rarely exceed a few microns in diameter, but which may grow to be several millimeters long. The interesting feature of these crystals is that they have a single dislocation, a spiral one running along the longitudinal axis. This is really a unique situation, since in the normal course of events both natural and artificially grown single crystals have a dislocation density of 10^6 to 10^8 per cm^3. The whiskers have values of critical shear stress approaching the values computed theoretically for perfect single crystals, elastic limits of over 10^6 psi, and elastic strains approaching 10 per cent.

8.10 Dislocation theory and slip

As we implied in the foregoing section, the observed plastic behavior of normal single crystals can largely be explained in terms of the motion of dislocations in a strain field. The mechanism of this motion was described in Sec. 4.3, but at that time we made no attempt to explain the mechanical consequences of the motion of dislocations.

The stress required to move an edge dislocation along a crystal plane is very much lower than that required to produce slip in a perfect crystal. This becomes evident when we consider the mechanism by which an edge dislocation travels through a crystal. None of the atoms moves very far from its initial position even though the dislocation goes all the way across the plane. Therefore, none of the bonds involved in the movement has to be stretched much further than the dislocation originally disturbed it. Also, only the bonds of atoms lying on the dislocation line are involved rather than those between all atoms on the plane.

When a dislocation has moved from its initial location to the surface of the crystal, it produces a step one lattice parameter deep at the surface and slip begins. When several dislocations on a single plane migrate to the surface, they will produce a step large enough to cause a stress concentration there. If this local stress is large enough, more dislocations will be generated on the same plane, and extensive slip will result. It is noteworthy that surface films, such as an adherent oxide coating, will raise the critical shear stress significantly. Possibly the reason is that the surface film will reinforce the crystal surface and tend to prevent dislocations from passing all the way to the sur-

face. This in turn would keep the local stress concentration from occurring, and would postpone slip until the external stress was raised to a higher level.

Since dislocation motion plays such a controlling role in the slip process, we may explain many aspects of plasticity in terms of factors that inhibit the motion of dislocations. In general, impurity atoms, grain boundaries, and other dislocations will decrease the mobility of a dislocation, while thermal agitation of the lattice will increase dislocation mobility. The next three sections are concerned with the mechanisms by which these factors operate, and with the correlation of these with certain well established experimental observations on the effect of thermal and mechanical history and grain size upon the elastic limit of materials.

8.11 Strain hardening

If an annealed single crystal is subjected to a stress cycle in which it is first stressed beyond its elastic limit, and then the stress is removed and reapplied, it will be found that upon the second application of the stress, the elastic limit is higher than it was on the first. This phenomenon is called *strain hardening* or *work hardening*. It is also observed in polycrystalline materials, and it is often utilized in the fabrication of materials when it is important to extend the elastic range as far as possible.

Strain hardening is usually explained in terms of dislocations interfering with the motion of one another. When a material is stressed, all but a relatively small number of the crystallographic planes will be subjected to some shear, and dislocations will move through the crystal in a variety of directions on intersecting planes. Each dislocation is the center of an inhomogeneous strain field, and when two dislocation lines traveling in skew directions meet, each will impede further motion of the other. When a stress is applied to a material not previously subjected to a plastic deformation, the dislocations will be randomly distributed throughout the crystal. After the first application of stress, a significant fraction of these will have become entangled in one another's strain fields. On subsequent applications of stress, the fraction of *mobile* dislocations will be considerably lower, and the material will deform elastically until the applied stress is large enough to drive the dislocations through one another's strain fields.

Work hardening can be relieved by annealing. This occurs because the thermal energy present in the crystal lattice is in the form of strain waves traveling randomly through the lattice. If one of these strain waves has sufficient amplitude at the instant that it passes a region in which two dislocations are hindered by their mutual strain fields, they may be able to pass through one another. The frequency of occurrence of such events depends upon the absolute temperature. In general, if a material has been work hardened, the elastic limit will decrease with time to the value that it had

before work hardening, the rate of decrease being dependent upon the absolute temperature. In some materials this will occur within a relatively short time even at room temperature. In others, the process is so slow that the material may be deliberately work hardened and design calculations can be based upon the new elastic limit even if the material is to be used for years at room temperature. The presence of alloying elements has a strong influence upon the rate at which work hardening is relieved by thermal agitation of the lattice.

8.12 Alloying

Even very small concentrations of impurities in a crystal lattice will raise the elastic limit significantly because of their ability to impede the motion of dislocations through the lattice. No two different kinds of atoms will have exactly the same bonding characteristics, and even pairs such as copper and silver, whose atoms are nearly of the same size and which are very similar chemically, will cause local strain fields if either is present as an impurity in the lattice of the other. These strain fields are capable of inhibiting the motion of dislocations, causing a corresponding increase of the elastic limit. The most striking example of the ability of a small amount of alloying material to raise the elastic limit of a material is the alloy formed between copper and beryllium. An alloy of 3% beryllium in copper will have a modulus of elasticity that is almost identical with that of pure copper, but the elastic limit is so much higher than that of copper that it would be possible to replace spring steel by the copper-beryllium alloy in many applications.

8.13 Polycrystalline materials

The grain boundaries of a polycrystalline material have a pronounced effect upon the plastic behavior of a solid. Any interface will generate a strain field that will have an appreciable magnitude for a distance of 10 to 20 Ångstrom units behind the interface. This strain field occurs because the atoms on the interface are not symmetrically bonded to neighbors across the interface. These strain fields contain stored energy, and the energy per unit volume of small crystals will be higher than that of large crystals. This has the effect of raising their chemical potential and they will be more active chemically than their neighbors. For the same reason, small crystals tend to disappear when the sample is annealed, while the large crystals tend to grow still larger at the expense of their small neighbors.

The strain fields at crystal interfaces are capable of inhibiting the motion of dislocations, and polycrystalline materials will normally be elastic Also, if the crystallographic axes are randomly oriented, a polycrystalline material will be macroscopically isotropic in all of its properties.

In addition to the inhibiting action of grain boundaries upon dislocation motion, the crystals tend to reinforce one another. The local stress in a small region of a polycrystalline sample may have a magnitude and orientation sufficient to produce slip in some of the crystals in the region but they will be restrained and reinforced by neighboring crystals whose crystallographic planes are more favorably oriented with respect to the applied stress.

In Chap. 4, we remarked that in the fabrication of materials, particularly metals, many factors in the fabrication process itself tend to introduce a preferred bias in orientation of the crystallographic axes of the individual crystals

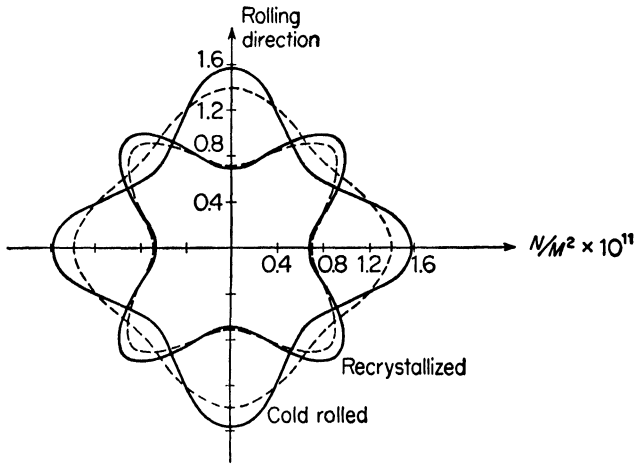


Figure 8.10 Anisotropy of Young's modulus in rolled copper sheet. (After Schmid and Boas, *Kristallplastizität.*)

in a polycrystalline material. This may cause the material to become anisotropic with respect to some or all of its properties. In the case of fcc crystals, the ⟨112⟩ or an equivalent crystallographic direction tends to align with the rolling direction, while bcc materials tend to have the ⟨110⟩ crystallographic direction become aligned with the rolling direction. If the material is subsequently annealed, we often find an entirely different bias developing. In fcc materials that have been cold rolled and annealed, the ⟨100⟩ crystallographic axis will have a bias in the rolling direction. For bcc materials, various investigators have found a bias in the rolling direction of the ⟨112⟩, ⟨110⟩, and ⟨100⟩ crystallographic directions, depending upon the annealing conditions. In other cases, annealing restores the original isotropy of the material. Figure 8.10 shows the directionality of Young's modulus for copper (fcc). The outer solid contour shows the directionality of Young's modulus for a single crystal measured upon a plane perpendicular to the ⟨110⟩ direction, and the inner solid contour shows the directionality of Young's modulus as measured upon a plane perpendicular to the ⟨100⟩ direction. The outer dashed contour shows

the same data for a cold rolled copper sheet and the inner dashed contour shows it for the same sheet after it has been annealed

The behavior of copper is typical of fcc metals; cold rolling tends to cause the ⟨110⟩ crystallographic direction to become aligned with the rolling direction. This will cause the elastic modulus of the material to become anisotropic even though it is polycrystalline. However, a certain fraction of the crystals will escape alignment, and the properties will be intermediate between those of single crystal copper and polycrystalline copper. When the sheet is annealed, apparently those crystals with their crystallographic axes aligned with the rolling direction grow at the expense of the other crystals, with the result that after annealing the polycrystalline material exhibits almost the same directionality of properties as would be observed from measurements on single crystals on a plane perpendicular to a ⟨100⟩ direction.

8.14 Creep

Heretofore we have ignored time as a factor in the stress-strain relationship. As a matter of fact, many materials will respond to a stress by showing a

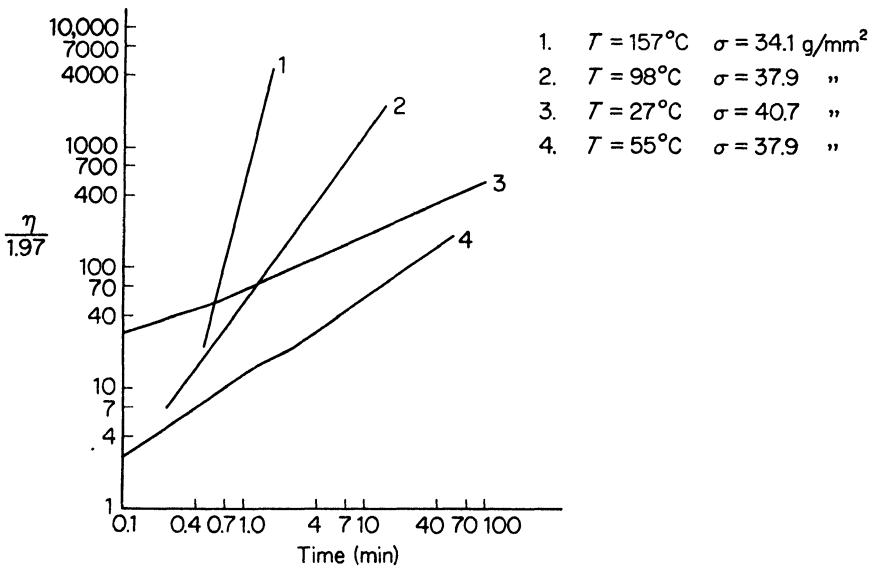


Figure 8.11(a) Creep rate of a single crystal Zn. η = shear resolved on plane of easiest slip. (D. Thompson, *J. App. Phys.* **26,** 280 (1955).)

prompt strain increment which subsequently increases slowly with time while the stress is held constant. This phenomenon is called *creep*. Practically all materials will creep at elevated temperatures. The steady-state creep rate is temperature dependent and the curves of Figs. 8.11a, 8.11b, and 8.11c pre-

sent data typical of many materials. These curves represent the creep behavior for a single stress level.

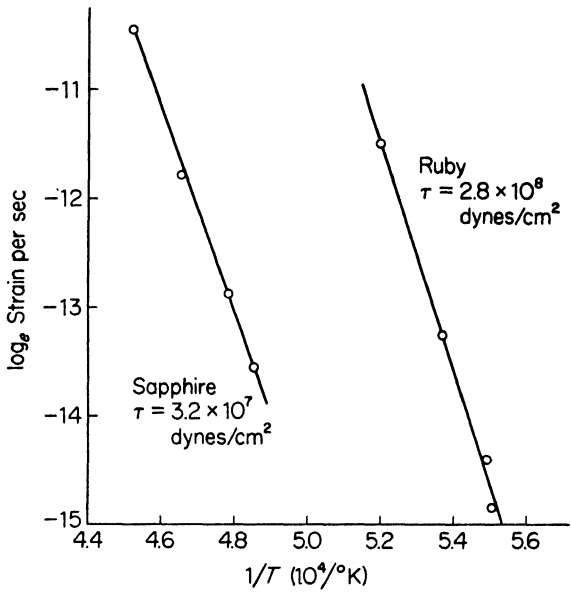


Figure 8.11(b) Steady state creep rate versus temperature of a ruby and sapphire. (R. Chang, *J. App. Phys.* **31,** 484 (1960).)

Generally the creep rate R and the absolute temperature will be related by an equation of the form

$$\frac{\partial \ln R}{\partial(1/T)} = \text{const.}$$

which will be valid at least in certain temperature ranges. In a later chapter we shall develop quasi-theoretical reasons for expecting this behavior. In part, creep can be explained by the fact that dislocations that have been entrapped by impurity sites or on other dislocations may receive enough energy from the thermal excitation of the lattice to surmount the strain barriers. The probability that this will happen increases exponentially with temperature.

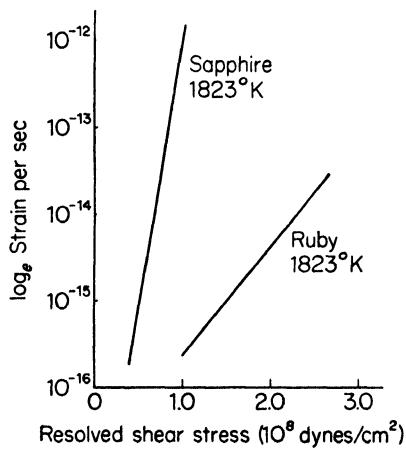


Figure 8.11(c) Steady state creep rate versus resolved shear stress for a ruby and a sapphire. (R. Chang, *J. App. Phys.*, op. cit.)

PROBLEMS

8.1 A square having sides of unit length is distorted so that one pair of sides is elongated by 10^{-3} distance units. The other pair of sides remain unchanged. The sides that were elongated also rotate clockwise through an angle of 0.01 radians. The other pair of sides maintain their original orientation in space. Using a coordinate system whose axes coincide with the sides of the square before distortion, evaluate the components of the rotation dyad **R** and the components of the strain tensor $\boldsymbol{\varepsilon}$.

8.2 If an isotropic material has the shape of the square in the previous problem and the stress required to produce the distortion of that problem has the components

$$\sigma_{11} = 10^9 \text{ dynes/cm}^2 \qquad \sigma_{12} = 5 \times 10^8 \text{ dynes/cm}^2$$

$$\sigma_{13} = \sigma_{23} = \sigma_{22} = \sigma_{33} = 0$$

calculate Y, G and μ.

8.3 From the results of Prob. 8.2, compute the nonzero elements of the elastic constant matrix and the nonzero components of the elastic rigidity tensor.

8.4 Show that crystal class 3 has thirteen nonzero independent elastic constants.

8.5 Using the data of Table 8.2, evaluate the technical elastic moduli for uranium, magnesium and nickel.

8.6 Combining the information in Tables 8.1 and 8.2, evaluate the nonzero components of the elastic rigidity tensor for uranium, magnesium, and nickel.

8.7 Using the results from Prob. 6, evaluate the stresses that would be required to produce the strain described in Prob. 8.1 in single crystal samples of uranium, magnesium and nickel.

8.8 Using highest areal density of atoms as the criterion for specifying a slip plane, and the highest linear density of atoms as the criterion for specifying the preferred slip direction, list the various combinations of crystallographic planes and directions that constitute the slip systems for uranium, nickel, and magnesium.

8.9 In terms of the conventional coordinate system of Table 4.1, the components of a stress in a particular system of units are

$$\sigma_{11} = 5 \qquad \sigma_{22} = 2 \qquad \sigma_{33} = 0 = \sigma_{13} = \sigma_{23} \qquad \sigma_{12} = 10$$

Find the resolved shear stress for the most sensitive member of the slip system of nickel, magnesium and uranium. "Most sensitive member" refers to the slip plane and direction sustaining the greatest shear stress.

8.10 The steady state creep rate of a material at 1000°K is 10^{-8} sec^{-1} and at 800°K is 10^{-10} sec^{-1} under the same stress. If asked for an informed opinion about the steady state creep rate at 700°K and denied access to any further information, what value would you give and what qualifications about the validity of this value would you offer?

9

THERMAL PROPERTIES OF SOLIDS

Thermal energy may be present in a solid in a variety of forms corresponding to the mode of motion of its fundamental particles. Among these we find vibration of the lattice points around an equilibrium position, rotation of functional groups in a molecule, and translation of electrons. Except in a few cases, the rotation of molecular groups does not contribute significantly to the total thermal energy, and electronic translation is only important in metals at elevated temperatures. Consequently if we consider only lattice vibrations, we shall have discussed the most important mode.

These lattice vibrations are directly involved in such typical properties as specific heat, thermal conductivity, and thermal expansion. In addition to these, however, we shall find that they are involved in the mobility of charge carriers in solids, in the concentration of charge carriers in semiconductors, in diffusion rates, in phase transformation rates, and in influencing the mobility of dislocations with all of their extensive mechanical consequences. We will not be able to derive quantitative relationships between thermal vibrations of the lattice and all of these phenomena, but even a qualitative consideration of their effects provides a unifying principle through which rates of change of unrelated phenomena may be correlated with temperature.

9.1 The concept of a phonon and the natural vibrations of a solid body

We cannot treat the vibration of a lattice point in a solid as though it were isolated and vibrating in its own potential well. Each lattice point is strongly

coupled to its neighbors through the bonding forces that hold the solid together. Thus, the thermal energy present as lattice vibrations assumes the form of strain waves which move through the lattice with the speed of sound. These strain waves are called *phonons*, and for many purposes they may be regarded as particles capable of colliding with other particles and transferring energy and momentum to the particles with which they interact. In many respects, phonons are identical with sound waves. They differ from sound waves in two important ways: (1) A sound wave may have an arbitrary frequency, whereas phonon frequencies are restricted to natural vibrational frequencies of the solid; (2) Sound waves are largely coherent, although some dispersion may be introduced by the solid through which they pass; phonons, on the other hand, may travel in almost any direction through the lattice independently of the other phonons that may be present.

In essence, a phonon is a transient lattice defect, and much of the action of phonon in modifying the properties of a solid may be understood by regarding them as such. Phonons may be scattered by impurities, grain boundaries, or any other source of lattice distortion. So in our attempt to find the spectral distribution of the phonons in a solid, we must idealize the solid and postulate that it is free of lattice defects, and that the phonons do not interfere with one another.

We remarked earlier that the phonons are restricted to wavelengths dictated by the external dimensions of the body. From this fact, we may derive their spectral distribution, and we shall begin by considering a rectangularly shaped solid. This might seem to be lacking in generality, but since our objective is to find the number of phonons in a given range of wavelengths per unit volume, we shall find that the external dimensions cancel out and leave us with a result that is applicable to a body of arbitrary shape.

Let a, b, and c represent the dimensions of the body. In order to correspond to a natural vibration of the body, a phonon traveling in a direction making the angle α with side a, the angle β with side b, and the angle γ with side c will be restricted to wavelengths

$$\frac{1}{\lambda}\cos\alpha = \frac{n_1}{2a} \qquad \text{(a)}$$

$$\frac{1}{\lambda}\cos\beta = \frac{n_2}{2b} \qquad \text{(b)} \qquad (9.1)$$

$$\frac{1}{\lambda}\cos\gamma = \frac{n_3}{2c} \qquad \text{(c)}$$

where the n's are integers. Equations (9.1) may be written

$$\left(\frac{1}{\lambda}\right)^2 = \left(\frac{n_1}{2a}\right)^2 + \left(\frac{n_2}{2b}\right)^2 + \left(\frac{n_3}{2c}\right)^2 \qquad (9.2)$$

This is identical with Eq. (6.15), which expresses the limitation imposed upon the wavelength of electrons by the external dimensions of the body. Consequently, we may use an n-space to deduce the spectral distribution of phonons in a body in the same manner as we deduced the density of the quantum states of electrons in Chap. 5. The identity of form carries through completely, including the restrictions upon the short-wave end of the phonon spectrum imposed by the interactions of the phonons with the lattice. It makes no sense to speak of a lattice vibration whose wavelength is less than twice the lattice spacing in the direction of propagation. Thus the limitation of the lattice upon the phonon spectrum is identical with the effect of the first Brillouin zone boundary upon the electron waves. The only difference is that it imposes an absolute limit upon the short-wave end of the phonon spectrum, while it only interrupts the spectrum of electron waves.

From these considerations we may write the spectral distribution function for phonons directly from Eq. (6.16)

$$dN\left(\frac{1}{\lambda}\right) = 4\pi\left(\frac{1}{\lambda}\right)^2 d\left(\frac{1}{\lambda}\right) \tag{9.3}$$

Again this equation is applicable only for values of $1/\lambda$ less than or equal to the radius of the sphere in n-space which is just tangent to the inside of the first Brillouin surface. There are higher values of $1/\lambda$ permitted, but only in specific directions. The total number of values which $1/\lambda$ may assume is finite, and corresponds to the number of points in n-space that satisfy Eq. (9.2) and lie entirely inside the first Brillouin surface. Thus if the spectral distribution of the phonons in a solid is plotted, as in Fig. 9.1, the solid portion of the curve is governed exactly by Eq. (9.3) until $1/\lambda$ has a value that will touch the inside of the Brillouin zone at some point. After this, the spectral distribution will vary from one crystal type to another and will be governed by the exact shape of the Brillouin zone. The dashed portion of the curve represents this part of the spectral distribution function.

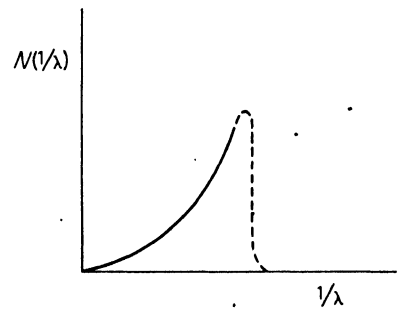


Figure 9.1 Spectral distribution of phonons in a solid.

9.2 The energy of a phonon and the internal energy of a solid

We have mentioned earlier that the classical physics of the later nineteenth century had difficulties in many areas when it came to correlating theory with

experiment. One of the tottering principles of that era was known as the *equipartition of energy*. According to this, the total energy of any material would be distributed among the particles in such a way that the average energy of each particle would be $\frac{1}{2}kT$ (k is the Boltzmann's constant) for each mode of motion. Thus a diatomic molecule would have $\frac{3}{2}kT$ for three translational degrees of freedom, plus $\frac{1}{2}kT$ for a vibrational mode, plus $\frac{1}{2}kT$ for a rotational mode. This would imply that the average vibrational energy of a particle of a solid would be $\frac{3}{2}kT$ corresponding to the three principal modes of vibration leading to a constant specific heat of $\frac{3}{2}Nk$ for a solid. Except in metals where the electronic contribution becomes important at high temperatures, most solids actually approach this value at elevated temperatures, and many solids, including metals, have a value near this for their room temperature specific heats. However, at low temperatures the principle fails completely.

The first attempt to account for the observed variation of specific heat with temperature was by Einstein. Using Planck's equation[1] relating the average energy of an oscillator to the absolute T and the frequency of oscillation

$$E = \frac{h\nu}{\exp h\nu/kT - 1} + h\nu \tag{9.4}$$

Einstein assumed that all of the particles in a solid oscillate at the same frequency and that the total internal energy is

$$U = N\bar{E}$$

and

$$C_v = \frac{\partial U}{\partial T}\bigg)_v$$

Although this approach gives a variation of C_v which is qualitatively correct, it makes no provision for difference between materials and fails completely at very low temperatures. Therefore we shall use a more modern approach based upon Planck's equation.

If we knew the spectral distribution of phonon frequencies we could calculate the total energy of a solid in the following manner: Let $dN(\nu)$ be the number of phonons per unit volume having frequencies between ν and $\nu + d\nu$. Then the total energy per unit volume would be

$$U_{\text{tot}} = \int_0^{\nu_{\max}} \left\{ \frac{h\nu}{\exp h\nu/kT - 1} + h\nu \right\} dN(\nu) \tag{9.5}$$

Equation (9.5) could be divided by an appropriate factor to reduce it to energy per mole or energy per gram. In this form it could be differentiated with respect to temperature to give the specific heat, or alternately it could be incorporated into any other thermodynamic relationship involving internal energy.

[1] This equation was based upon the postulate that the energy of the oscillation is quantized. It was validated by Planck's experiment with the radiation from a black body.

In the previous section we developed an expression for the spectral distribution of phonons in terms of their reciprocal wavelength. If we can find a valid way of converting this into a spectral distribution of frequencies, then we can develop the potentialities of Eq. (9.5). We shall do this presently, but the method we shall use involves certain approximations and assumptions which should be understood thoroughly before the development will make much sense.

The relationship between wavelength and frequency is

$$\lambda\nu = c$$

where c is the propagation velocity of the wave. If there were a single velocity of sound in a solid, our task would be very easy, but even in isotropic solids, transverse and longitudinal waves are propagated at different velocities, and in anisotropic solids there will be as many distinct propagation velocities as there are distinct technical elastic moduli. Consequently, one problem in working with Eq. (9.5) is how to relate the spectral distribution of the frequencies to the spectral distribution of the reciprocal wavelengths. Another problem occurring at this point is how to provide for that part of the spectral distribution function which arises from those permitted values of $1/\lambda$ which lie between the Brillouin surface and the largest sphere which can be included inside the Brillouin surface.

To handle these problems, we assume that the Brillouin zone is spherical, and that the solid under consideration is isotropic. For longitudinal waves only

$$\lambda\nu = c_l$$

Substitution into Eq. (9.3) gives the spectral distribution of longitudinal waves in frequency

$$dN(\nu) = \frac{4\pi\nu^2}{c^3}\,d\nu \tag{9.6}$$

The portion of the lattice energy per unit volume that is due only to longitudinal waves is

$$U_l = \frac{4\pi}{c_l^3}\int_0^{\nu_{\max(\text{longitudinal})}} \left\{\frac{h\nu^3}{\exp h\nu/kT - 1} + h\nu^3\right\} d\nu \tag{9.7}$$

Transverse waves can be circularly polarized, and consist of two mutually perpendicular components. Each of these would carry the same amount of energy, and by analogy with (9.7) we can express that portion of the lattice energy due to transverse waves by

$$U_t = \frac{8\pi}{c_t^3}\int_0^{\nu_{\max(\text{transverse})}} \left\{\frac{h\nu^3}{\exp h\nu/kT - 1} + h\nu^3\right\} d\nu \tag{9.8}$$

The total internal energy then would be the sum of Eqs. (9.7) and (9.8).

Peter Debye in 1912 developed an approach which is still employed. He

used a weighted average of both wave types for the velocity of propagation which is defined by the equation

$$\frac{3}{c^3} = \frac{2}{c_t^3} + \frac{1}{c_l^3}$$

He assumed that both types of wave would have the same maximum frequency. This results in

$$U = \frac{3v}{c^3} \cdot 4\pi \int_0^{\nu_{max}} \left\{ \frac{h\nu^3}{\exp h\nu/kT - 1} + h\nu^3 \right\} d\nu \tag{9.9}$$

When Eq. (9.9) is differentiated with respect to temperature, the derived values of specific heat are in close agreement with observed values of specific

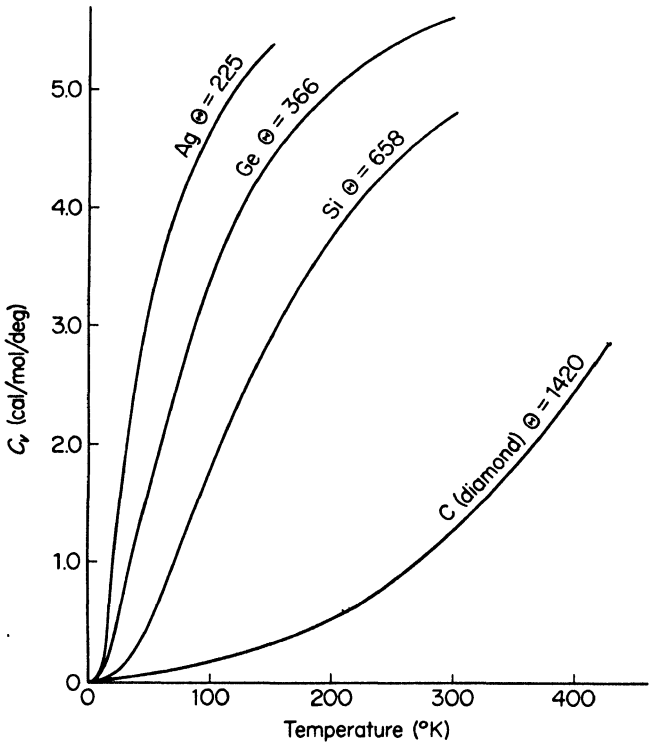


Figure 9.2 Variation of specific heat at constant pressure for materials having different characteristic temperatures.

heat. The method seems to apply also to anisotropic materials. Of course, the integrals in (9.7) through (9.9) have to be evaluated numerically and there are two methods for selecting the value of ν_{max} for the upper limit of the integration. One involves using ν_{max} as an adjustable parameter which can be evaluated from a single known value of specific heat of a given material at any one temperature. The parameter then fixes the specific heat curve for

that material at all temperatures and in this form it may be checked experimentally. As we remarked before, correlation with experiment is surprisingly good. Usually, instead of specifying a value of ν_{max} for a material, a quantity

$$\Theta = \frac{h\nu_{max}}{k}$$

called the *characteristic* or *Debye temperature* is specified.

Alternatively, a value of the characteristic temperature may be derived directly from the elastic constants of an isotropic material by combining the equations

$$\frac{3}{c^3} = \frac{1}{c_l^3} + \frac{2}{c_t^3}, \qquad c_l = \sqrt{\frac{Y}{\rho}}$$

$$c_t = \sqrt{\frac{G}{\rho}}, \qquad G \simeq \frac{3}{8} Y$$

to produce

$$c = 0.7 \sqrt{\frac{k}{\rho}} \tag{9.10}$$

Since the volume of the Brillouin zone in n-space is equal to the number of atoms per unit volume in Euclidean space, we may write

$$\frac{N}{V} = \frac{4\pi}{3} \left(\frac{1}{\lambda_{min}} \right)^3$$

where $N = 6.023 \times 10^{23}$, and V is the molar volume.

Because $1/\lambda_{min} = \nu_{max}/c$, we have from Eq. (9.10)

$$\nu_{max} = 0.7 \sqrt{\frac{k}{\rho}} \left(\frac{3}{4\pi} \frac{N}{V} \right)^{1/3}$$

Notice that Θ is related to λ_{max} which in turn is governed by the elastic constants. These are temperature dependent; consequently Θ is also a function of temperature.

Figure 9.2 shows specific heat versus temperature curves for several materials, and Table 9.1 lists some characteristic temperatures.

Table 9.1 *Characteristic temperatures of certain materials**

Material	Θ (°K)
Be	1160
Mg	406
Fe	467
Au	165
Tl	89
NaCl	308
KCl	230
Ag	225
Zn	308

* Reprinted with permission from Charles Kittel, *Introduction to Solid State Physics* (2nd ed.; New York: John Wiley & Sons, Inc., 1956).

The degree of agreement between the theoretical treatment of the variation of specific heat presented here and those which are observed experimentally in anisotropic single crystals is fairly good, but even this may be improved by using different characteristic temperatures for longitudinal and transverse waves, and for waves of both types propagated in different directions through the crystal.

9.3 Thermal expansion

The thermal expansion of solids may be attributed to the increase in average amplitude of oscillation of the lattice points as the thermal energy increases.

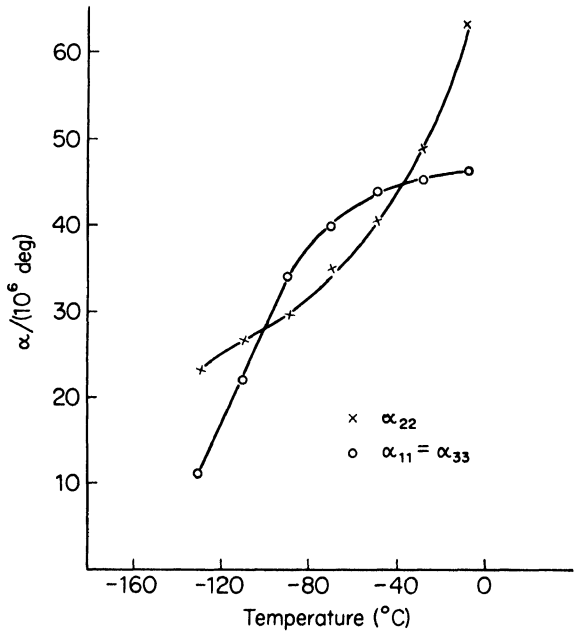


Figure 9.3 Temperature dependence of the components of the thermal expansion tensor of ice. (S. La Placa and B. Post, *Acta Cryst.* **13**, 503 (1960).)

From a purely phenomenological point of view, the thermal expansion of a solid may be completely characterized by a rank two tensor. This is a symmetric tensor and is subject to the usual symmetry restrictions discussed in Chap. 7. Thus materials having a cubic structure are isotropic in their thermal expansion, those having tetragonal or trigonal symmetry have two independent thermal expansion coefficients, and those having a lower symmetry have three independent thermal expansion coefficients. The set of differential equations from which the strain due to a temperature change

would be computed is

$$d\epsilon_{ij} = \alpha_{ij}\, dT$$

Over relatively small temperature ranges the components of $\boldsymbol{\alpha}$ usually may be considered constant. Over larger temperature ranges the components of $\boldsymbol{\alpha}$ vary in a manner similar to the variation of C_v with temperature discussed in the previous section. There are quasi-theoretical reasons for this, but perhaps these should not be taken too seriously since there are materials in which some of the components of $\boldsymbol{\alpha}$ are actually negative in certain temperature ranges.

Figure 9.3 shows the variation of both components of the thermal expansion tensor of ice with temperature. Table 9.2 lists the room temperature thermal expansion coefficients of assorted materials.

Table 9.2 *Thermal expansion coefficients of single crystal solids at room temperature*

Material	$\alpha_{11} \times 10^6$	$\alpha_{22} \times 10^6$
Bi	11.7	17.4
Mg	25.0	27.0
Cd	17.0	49.0
SiO_2	9.0	14.0
$CaCO_3$	−6.0	25.0
Zn	14.0	63.0

9.4 Thermal conductivity

In terms of the systematic relationships developed in Chap. 6, the entropy current density at a point in a material is given by the equation

$$\mathbf{J}_s = -\mathbf{C}_s \cdot \nabla T - \mathbf{C}_{se} \cdot \nabla V$$

where the **C**'s are conductivities. In the present section we do not wish to consider the cross effects, and we shall deal only with the cases in which the electric field intensity ∇V is zero. The conductivity $\mathbf{C}_s$ is a rank two tensor property of the material under consideration and again is subject to the restrictions imposed by Neumann's principle.

The steady-state temperature distribution in a material may be deduced from solutions of

$$\nabla \cdot \mathbf{J}_s = -\nabla \cdot \mathbf{C}_s \cdot \nabla T = 0 \tag{9.11}$$

Transient systems are governed by continuity considerations, and equating the net outflow from a region to the rate of entropy production in that region

$$\int_\Sigma \mathbf{J} \cdot d\boldsymbol{\sigma} = -\int_\tau \rho \frac{\partial S}{\partial t}\, d\tau$$

followed by the application of divergence theorem

$$\int_{\Sigma} \mathbf{J} \cdot d\delta = \int_{\tau} \nabla \cdot \mathbf{J}\, d\tau$$

produces the transient equation

$$\nabla \cdot \mathbf{J} = -\rho \frac{\partial S}{\partial T} = -\nabla \cdot \mathbf{C}_s \cdot \nabla T \tag{9.12}$$

These relationships are rigorous and do not depend upon the mechanisms of conduction. Consequently the solutions of Eqs. (9.11) and (9.12) will suffice to describe either the transient or steady-state temperature distribution in a body for a given set of boundary conditions so long as the components of the conductivity tensor $\mathbf{C}_s$ are known in some coordinate system.

Just as the technical elastic moduli are used in preference to the elastic constants in representing results, we have here a situation where there is a difference between the parameters which may be interrelated systematically and those which are convenient to measure. Normally, the thermal conductivity is not determined in the form of an evaluation of the components of the tensor $\mathbf{C}_s$ but in terms of the power transmitted per unit area in the direction of the temperature gradient. In this scheme the thermal conductivity is defined by the equation

$$w_1 = -K \frac{\partial T}{\partial x_1} \tag{9.13}$$

where w_1 is the power per unit area transmitted in the x_1 direction.

The difficulty with Eq. (9.13) is that in anisotropic materials, the components of the temperature gradient in the x_2 and x_3 directions can contribute to the heat flow in the x_1 direction. In contrast with this, expansion of

$$\mathbf{J}_s = -\mathbf{C}_s \cdot \nabla T$$

produces[2]

$${}_sJ_1 = {}_sC_{11} \frac{\partial T}{\partial x_1} + {}_sC_{12} \frac{\partial T}{\partial x_2} + {}_sC_{13} \frac{\partial T}{\partial x_3}$$

This can be applied directly and without further qualification to any material, whereas (9.13) can only be applied as it stands to cubic or isotropic materials. We may, however, define the set

$$\begin{aligned} w_1 &= -K_1 \frac{\partial T}{\partial x_1} \qquad \left(\frac{\partial T}{\partial x_2} = \frac{\partial T}{x_3} = 0\right) \\ w_2 &= -K_2 \frac{\partial T}{\partial x_2} \qquad \left(\frac{\partial T}{\partial x_1} = \frac{\partial T}{\partial x_3} = 0\right) \\ w_3 &= -K_3 \frac{\partial T}{\partial x_3} \qquad \left(\frac{\partial T}{\partial x_2} = \frac{\partial T}{\partial x_1} = 0\right) \end{aligned} \tag{9.14}$$

[2] The subscript s appears before J_i and C_{ij} to permit us to denote a particular component of either with less confusion.

These could be applied to crystals having orthorhombic or higher symmetry, provided that we also apply the restriction that the temperature gradient must coincide with the crystallographic axes. Then

$$\frac{K_1}{T} = {}_sC_{11}, \qquad \frac{K_2}{T} = {}_sC_{22}, \qquad \frac{K_3}{T} = {}_sC_{33} \tag{9.15}$$

These relationships are entirely phenomenological. The actual mechanism by which energy is transported from a region of higher temperature to one of lower temperature involves a transport of phonons and/or electrons. The relative importance of these depends upon the material. In metals, about 90 per cent of the energy transport is effected by electrons. In semiconductors, the concentration of electrons available for energy transport depends very strongly upon temperature. When they are available, electrons are more effective than phonons in the transport of energy. In insulators, phonon transport is the only mechanism operative. It is this category that we shall consider here, deferring a treatment of semiconductors and metals to a later section.

It will be recalled that phonons are scattered by any lattice defect as well as by other phonons. The actual mobility of a phonon depends upon its velocity of propagation and the mean free path, while the driving force that would cause a net flow of phonons in a particular direction would be an enthalpy gradient. Thus

$$w = -\frac{\partial H}{\partial x} lc = -\rho C_p lc \frac{\partial T}{\partial s} \tag{9.16}$$

where H is the enthalpy, l the mean free path, and c the velocity of propagation of the phonons in the direction of the temperature gradient. The other symbols have their usual meanings. From Eq. (9.13), it is apparent that

$$K = \rho C_p lc \tag{9.17}$$

Recognizing that the phonons are scattered both by lattice defects and by other phonons, we can separate l into two parts

$$\frac{1}{l} = \frac{1}{l_s} + \frac{1}{l_t} \tag{9.18}$$

where $1/l_s$ is the scattering factor due to permanent lattice defects and $1/l_t$ is the phonon-phonon scattering factor.

At ordinary temperatures, most of the energy is associated with short-wavelength phonons and the structural scattering factor is almost independent of temperature. At very low temperatures, the longer-wavelength phonons begin to carry an appreciable proportion of the thermal energy and these are not scattered so readily by permanent lattice defects. Consequently, we may regard $1/l_s$ to be constant over a wide range of temperatures. Below

about 100°K, $1/l_s$ will begin to decrease with decreasing temperature and around 20°K, the structural scattering begins to be determined by the external dimensions of the sample.

The thermal scattering factor $1/l_t$ is found experimentally to be almost proportional to the absolute temperature. We would, of course, expect thermal scattering to increase with increasing temperature. On the other

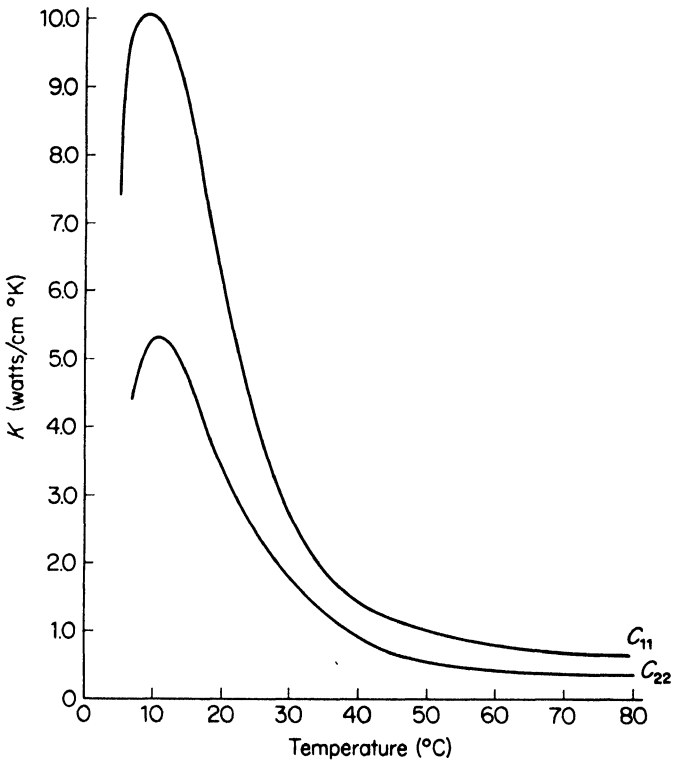


Figure 9.4 Temperature dependence of the components of the thermal conductivity tensor of single crystal quartz at low temperature.

hand, the structural scattering factor does not change much with temperature. Instead, it increases with increasing disorder of the solid. Thus in comparing samples of the same chemical species, we would expect a progressive increase in $1/l_s$ in the sequence: perfect single crystal-normal single crystal-polycrystalline sample-amorphous sample.

In addition to these factors affecting the mean free path, we must remember that C_p will influence K in Eq. (9.17) and c will give it the directionality that we would have expected from Eq. (9.14). Technically we should distinguish between the velocities of transverse and longitudinal

phonons, but as used here, c may be considered the weighted average of the velocities of propagation of both. Then, near absolute zero, we expect K to be very low because C_p approaches zero. As the temperature increases, C_p would increase as T^3. Thus K increases initially. However, in the range 10°K to 40°K, thermal scattering begins to cause a subsequent decrease in K.

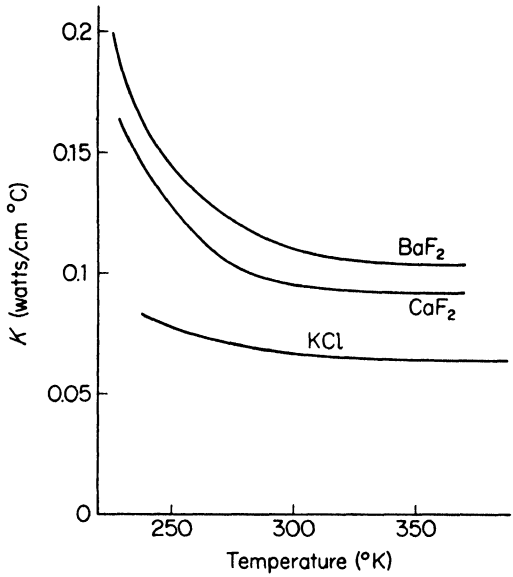


Figure 9.5 Temperature dependence of thermal conductivity of single crystal halides. (McCarthy and Ballard, *J. App. Phys.* **31,** 1410 (1960).)

Figure 9.4 illustrates this behavior in quartz. At higher temperatures, C_p becomes nearly constant and the decrease in l_t causes the continued decrease in K with increasing temperatures that is illustrated in Fig. 9.5. Amorphous materials behave quite differently from this. In these cases, structural scattering is much greater than thermal scattering and is nearly independent of temperature. Consequently, the variation of C_p with temperature is the governing factor.

9.5 Thermal conductivity of metals

Electrons are capable of transporting thermal energy; and at temperatures well above the Debye characteristic temperature, electrons in a metal transport most of the thermal energy. Consequently, one expects some relationship to exist between the electrical conductivity and the thermal conductivity of a metal. We develop this relationship in this section.

According to the free electron theory of metals,[3] the electrical conductivity of a metal is

$$C_{ev} = \frac{Ne^2\tau}{m} \tag{9.19}$$

where τ is the relaxation time, N the concentration of electrons, e their charge, and m their mass. The thermal conductivity due to the electrons is

$$K = \frac{\pi^2 kTN\tau}{3m} \tag{9.20}$$

where k is the Boltzmann constant. The ratio of the electrical conductivity

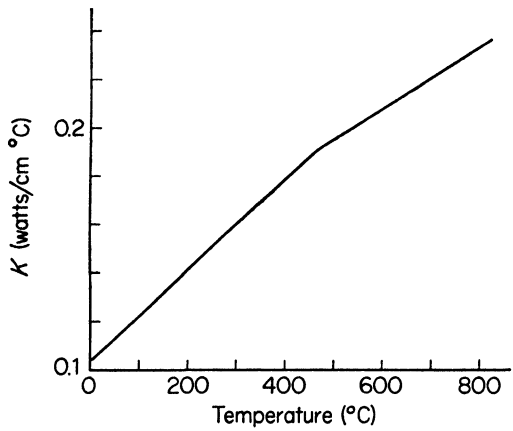


Figure 9.6 Thermal conductivity of Macloy G steel. (Powell and Tye, *Br. J. App. Phys.* **11**, (1960).)

to that portion of thermal conduction due to electron motion should be approximately the same for all metals at a particular temperature. We may remove the temperature from consideration by defining a quantity L by the equation

$$L = \frac{K}{TC_{ev}} = \frac{\pi^2 k^2}{e^2} = 2.45 \times 10^{-8} \text{ watt-ohms/deg}^2 \tag{9.21}$$

L is called the *Lorentz number*. Representative experimental values of the

[3] The free electron theory is at least as notable for its failures as for its successes. We shall not develop it in the chapter dealing with the electronic structure of solids. The interested reader is referred to Charles Kittel, *Introduction to Solid State Physics*, (2nd ed.; New York: John Wiley & Sons, Inc, 1956), or Adrianus K. Dekker, *Solid State Physics* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1957).

Lorentz number of several metals are given in Table 9.3. Figure 9.6 illus-

Table 9.3 *The Lorentz Number of Some Materials*

Material	$L \times 10^8$ *watt-ohm deg*$^{-2}$
Ag	2.31
Au	2.35
Cu	2.23
Mo	2.61
Sn	2.52
Zn	2.31

* From Kittel, *Introduction to Solid State Physics.*

trates the nearly linear increase in thermal conductivity predicted by the free electron theory.

9.6 Rate processes and activation energy

Many physical processes take place at a finite rate which is temperature dependent. Furthermore, we often find that the rates of diverse processes will vary with temperature in a similar manner in certain temperature ranges. To be more specific, we frequently find that the logarithm of the process rate will be linear with the reciprocal of the absolute temperature over wide ranges of temperature. This may be true of the rate of conduction of charge under a unit voltage gradient in a semiconductor, the rate of diffusion of one substance through another under a unit concentration gradient, the rates of certain chemical reactions such as esterification of an organic acid with an alcohol, the rate of failure of certain types of capacitors under life test conditions, and the rate of creep in metals under a load. This is a partial list, and with very little effort the number of common examples could be doubled or tripled. In every case that behaves in this manner there will be two common factors: there will be a tendency for the system to undergo a transformation from a higher energy state to a lower one, and this transformation will occur only when a critical member of the system gains an increment of energy above its average energy. We can illustrate this by using a variation upon an old pedagogical chestnut. Let us imagine that there are a large number of wooden blocks standing on edge on a flat surface as shown in Fig. 9.7a. Each of these has the potentiality of falling over on its side, since the center of gravity would thereby be lowered and hence the potential energy of each would decrease. However, if we look at the mechanics of the transformation from the state of standing on edge to lying on a side, we find that as a block tips up on a corner, the potential energy must first rise through a maximum, and then as the tipping progresses, it will decrease to its final value. This is characteristic of all of the phenomena in the class we are now discussing. In more general terms, if we follow the energy of each block as it undergoes the transformation, the total energy will change in the manner

shown in Fig. 9.7b. The initial energy increment is called the *activation energy* symbolized by E. Now we can carry this mechanical analogy one step further. Let us imagine that the blocks of wood are spaced so that they do not knock each other over as they fall down, and let us imagine that they are standing on a rubber diaphragm. If we introduce mechanical disturbances of random amplitudes at several points in the diaphragm, the resulting waves will interfere constructively at some points and destructively at others. These waves might be considered to be analogous to the phonons in a lattice. Now the blocks will fall over at a rate that is proportional to the concentration of diaphragm waves having energies greater than the energy necessary to tip a block up on an edge.

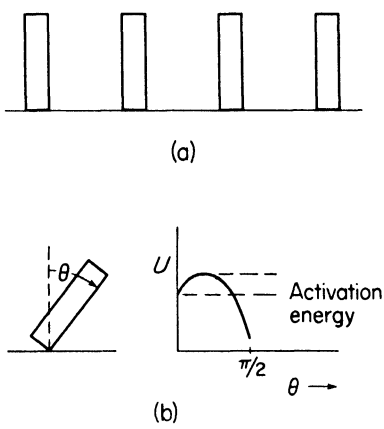


Figure 9.7 Mechanical model of a system in a metastable state from which a transition to the stable state would involve an initial increase in energy.

We might as well exploit this analogue to its limit. If the blocks were spaced sufficiently close to one another that they would strike one another as they fell, then the rate would no longer be controlled by the waves in the rubber diaphragm. It would only be necessary to have a few sufficiently energetic diaphragm waves to knock over one or two blocks, which would in turn knock over the rest of the blocks in short order. This is analogous to what happens if you test for the presence of illuminating gas by lighting a match, and it represents a class of processes different from that under discussion in this section.

The analogy cited in the preceding paragraphs is typical of the class of transformations that are called *activated processes*. Not all physical processes fall into category, but the theory of activated processes is more thoroughly developed than the theory of other types of transformations such as the chain reaction type. In general terms, activated processes proceed by some mechanism which at some point requires thermal energy to make it operative. The rate at which it will occur depends upon the concentration of phonons in the solid having an energy in excess of the activation energy. It is possible to derive theoretical equations which relate the absolute temperature to the concentration of phonons having energies in excess of the activation energy. This discipline is known as *statistical mechanics*, and these theoretical relationships are often based upon postulates selected to represent the truth as closely as possible without introducing too many mathematical difficulties. The exact form of the theoretical relationships derived from statistical mechanics may vary with the postulates used, but they always have the

common factor $e^{-E/kT}$. Thus from a theoretical point of view we might expect that activated processes will take place at a rate R such that

$$\frac{d \ln R}{d(1/T)} = -\frac{E}{k} + \text{other terms}$$

Frequently the "other" terms are negligible, and experimentally it is found that for activated processes a plot of the logarithm of the rate versus the reciprocal of the absolute temperature is linear over a wide range of temperature. The negative of the slope of this experimental curve is interpreted as the activation energy divided by the Boltzmann constant. In some cases there may be multiple activation energies, each of which has a predominant effect in a certain temperature range. For example, the rate curve shown in Fig. 9.8 is typical of a case where a process can proceed by two different mechanisms having different activation energies. At low temperatures the effect of the low activation energy process predominates, and at high temperatures the high activation energy mechanism predominates. The classical example of such a situation is the diffusion of an impurity through a polycrystalline solid. Diffusion along the grain boundaries would normally have a much lower activation energy than diffusion directly through the crystals. The creep curves shown on p. 185 are typical examples of an activated process.

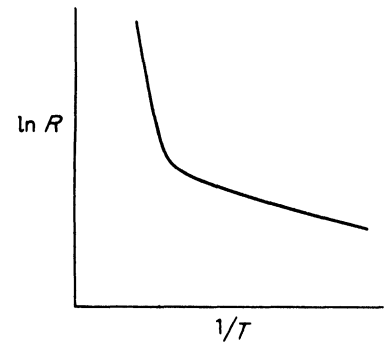


Figure 9.8 Curve showing variation of rate with temperature of a hypothetical rate process having two activation energies.

One note of caution should be sounded here. Just because diffusion, creep, or any of the other phenomena discussed here has been found to be an activated process in some materials does not mean that it will be an activated process in all materials. For example, the diffusion of copper through lead takes place in accordance with the principles discussed here. The diffusion of copper into germanium progresses rapidly to a low limiting concentration, but the rate at which this occurs is practically independent of temperature.

9.7 Metastability

Many common materials such as diamonds, hardened steels, and artificial fibers are actually unstable at room temperature with respect to a transformation into something else. However, they do not undergo such transformations, or at least do not undergo them at a measurable rate, and we use them every day with perfect confidence. When a system is capable of remaining in-

definitely in a state whose energy is higher than the state having the lowest possible energy, it is said to be *metastable*. We can explain metastability in terms of the activation energy. Consider a system that can be in either of two forms which we label A and B. Let A be the stable form above some temperature T, and B the stable form below T. At T the two forms may be equilibrium. The transformations from A to B and vice versa would be accompanied

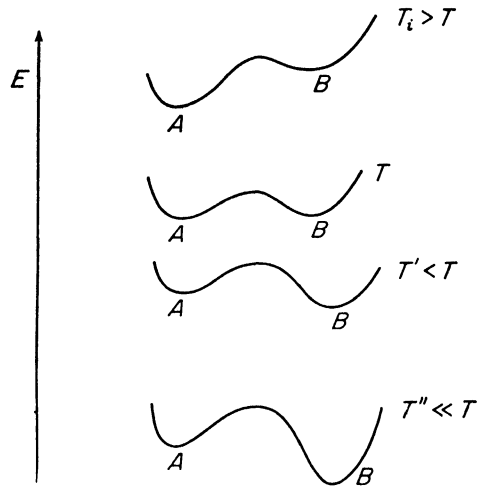


Figure 9.9 Schematic of the energy changes occurring in a system undergoing transitions between states A and B. In the upper curve, if the system were in state B, it would be unstable with respect to a transition into state A, but would have to surmount an activation energy barrier before the transition would occur. Just the reverse situation is illustrated in the bottom curve. In the second curve from the top, neither state is preferred on the basis of energy, and presumably, an assemblage of such systems would exist with a fraction in state A, and the rest in state B. These would be in equilibrium with each other, and the rate of transformation in either direction would be the same.

by the energy changes shown in Fig. 9.9. Now if the system is initially in state A at a temperature above T and is cooled to a temperature T' just below T, it will begin to transform into B at a rate which is determined by $e^{-E/kT'}$. If, instead, the system is cooled suddenly to a temperature T'' such that $kT'' \ll E$, then the transformation rate will become so slow that for practical purposes, the system will remain indefinitely in state A even though it is unstable with respect to transformation into B.

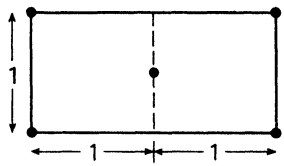
Technically even polycrystalline materials are unstable with respect to transformation into single crystals. Practically, though, the differences in energy are so small that the transformation is fairly difficult to achieve. However, if a polycrystalline material is annealed at some temperature below the

melting point, the larger crystals will grow at the expense of the smaller crystals.

Glasses are, of course, metastable with respect to transformation into the crystalline state, and glasses that are annealed at too high a temperature will begin to crystallize.

PROBLEMS

9.1 Construct the Brillouin surface for the two dimensional crystal whose basic cell is shown below:



9.2 Discuss the refinements that might be necessary to bring the results of Eq. (9.9) into closer agreement with theory for a tetragonal structure.

9.3 In a coordinate system chosen by the convention of Table 4.1 the components of the thermal expansion tensor of a crystal of class 3 are observed to be

$$\alpha_{11} = 6 \times 10^{-6} \qquad \alpha_{33} = 4 \times 10^{-6}$$
$$\alpha_{22} = 8 \times 10^{-6} \qquad \alpha_{23} = 4 \times 10^{-6}$$

(a) Compute the strains resulting from a 50° temperature rise.
(b) If the crystal were cut in the form of a cube, $\frac{1}{2}$ in. on a side and the coordinate axes coincident with the edge of the cube, describe the change in shape that would result from a 100° temperature rise.
(c) Find the orientation of the principal axes of the α tensor relative to the crystallographic axes.

9.4 At 20°C. the thermal conductivity of cadmium is 0.20 calories per sec per degree C per centimeter in a direction parallel to the hexagonal axis and in a direction perpendicular to the hexagonal axis the thermal conductivity is 0.25 in the same units. If the entropy current density vector has the magnitude of 0.01 entropy units (calories per degree) per square centimeter per second and makes an angle of 30° with the hexagonal axis, evaluate the components of the temperature gradient.

9.5 Assuming that diffusivity is an activated process, that the activation energy is constant and that only one activation energy is operative in the temperature range specified, calculate the activation energy for the diffusion process if $D = 3 \times 10^{-12}$ cm^2 sec^{-1} at 0°C and $D = 5 \times 10^{-10}$ cm^2 sec^{-1} at 100°C.

10

ELECTRONIC PROPERTIES OF SOLIDS

As a part of the treatment of bonding in solids presented in Chap. 5, the elementary zone theory was developed. This led logically to the existence of the electrons in bands of closely spaced states and permitted us to compute the density of states as a function of energy. Using these ideas, we were able to account qualitatively for the variation of electrical conductivity with valence for metals and to provide a theoretical base for the Hume-Rothery rules of alloying. Despite these accomplishments, the reader must have been aware that the entire theory was based upon a highly simplified mathematical model of the potential energy distribution in a solid. Because of this, many of the concepts have to be revised and stated in a more sophisticated manner to account for certain features of electrical conductivity, thermoelectricity, and other phenomena involving the motion of charge carriers.

We shall not attempt to base these revisions upon first principles but we shall strive for plausibility. The first sections of this chapter will be concerned with the aforementioned revisions and with the development of some additional ideas and concepts; the subsequent sections will be concerned with their applications.

10.1 The density of states function and the effective mass of charge carriers

It will be recalled that in the development of the elementary zone theory, the positions of the electrons in the solid were described in terms of waves whose

lengths were restricted to the values

$$\left(\frac{1}{\lambda}\right)^2 = \left(\frac{n_1}{2a}\right)^2 + \left(\frac{n_2}{2b}\right)^2 + \left(\frac{n_3}{2c}\right)^2 \tag{10.1}$$

by the requirement that the wave function be zero at the surface of the solid. This was combined with the de Broglie equation

$$p = \frac{h}{\lambda} \tag{10.2}$$

to impose restrictions upon the permitted values of momentum. From this the Brillouin zone was introduced to delineate the limits imposed by lattice diffraction. Although we did not state it at the time, it is evident that the loci of permitted values of momenta or of reciprocal wavelengths having the same energy fall on a sphere in p-space.

Using the equation

$$\frac{dE}{dp} = \frac{p}{m} \tag{10.3}$$

we were able to show how these conditions imposed a corresponding quantization upon the energy and thus deduced the density of states as a function of energy in the form

$$dN\,(E) = \frac{2\pi}{h^3}\,(2m)^{3/2}E^{1/2}\,dE \tag{10.4}$$

This equation is valid only for energies less than or equal to that value corresponding to a sphere tangent to two or more faces on the inside of the Brillouin surface.

This development of the zone theory must be revised because the velocity of propagation of the electron waves inside a solid, just as for light, is different from its free space value. Also, within a given crystalline solid it will vary with direction and wavelength. We may provide for this and still preserve the accustomed mathematical form by inventing a new quantity called *crystal momentum*. This has the dimensions of momentum, but differs in magnitude from true momentum by whatever amount is necessary to validate the de Broglie equation. Thus crystal momentum P, is related to wavelength by

$$P = \frac{h}{\lambda} \tag{10.5}$$

Naturally we would not expect Eq. (10.3) to provide the correct relationship between crystal momentum and energy. However, the equivalent of (10.3)

$$\frac{d^2E}{dp^2} = \frac{1}{m} \tag{10.6}$$

is a basic equation of both classical and quantum mechanics, and as such, we wish to preserve its form too. We may do this by substituting crystal momentum for true momentum and by defining a quantity, m^*, which we call the *effective mass*, such that

$$\frac{d^2E}{dP^2} = \frac{1}{m^*} \tag{10.7}$$

is valid.

The difference between the mass of a free electron and the effective mass of an electron traveling through a crystal is a result of the Bragg reflection by the lattice. If we define λ_B as the electron wavelength at which the lattice would cause Bragg reflection, then for $\lambda \gg \lambda_B$ the reflected component of the electron wave would be very low. To an observer, the electron would seem to respond to a force very much as would a free electron. As $\lambda \longrightarrow \lambda_B$, a successively larger component of the electron wave is reflected. In classical terms, a larger force would be required to give it a unit acceleration and the effective mass would be greater. When $\lambda = \lambda_B$, the electron wave would become a standing wave. It could not be accelerated and would have an infinite effective mass. When the reflected component is larger than the transmitted component of the wave, it would seem to an observer that the direction of the acceleration of the electron is opposite to the direction of the applied force, i.e. the effective mass would be negative.

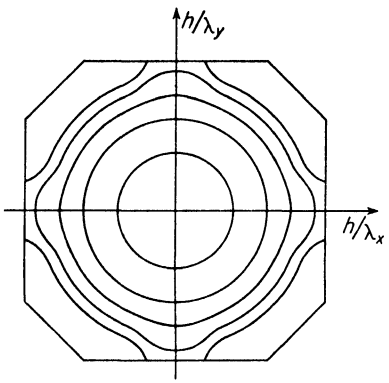


Figure 10.1 Plot of surfaces of equal energy in *P*-space.

The result of this is that a plot of E versus P would not have the parabolic form required by Eq. (10.3), but when $1/\lambda$ approaches $1/\lambda_B$, the energy surface would be distorted toward the zone boundary as shown in Fig. 10.1. The negative curvature corresponds to a negative effective mass.

Let us consider the consequence of this in a nearly empty band, a partially filled band, and a nearly filled band. If all of the electrons were in the lowest possible states, they would be paired in states corresponding to equal energy but opposite momenta. The situation is shown in Figs. 10.2a, 10.2b, and 10.2c by the solid circles. Figures 10.2a and 10.2b show a partially filled and a nearly empty band. If an electric field were present and directed so that it would exert a force to the right, electrons would leave the states symbolized ⊙ and enter states symbolized by ×, resulting in a net flow of electrons to the right. In the case of Fig. 10.2c, the band is nearly full. If the electron marked

⊙ were subjected to a force acting to the right, the reflected component of the wave would be so large that it would enter the state marked ×, leaving a vacancy in the position marked ⊙. This vacancy is called a *hole*, and it acts exactly as though it were a charge carrier having an effective mass as well as all of the other attributes of an electron except that its charge is positive. The result is that when the effective masses of electrons are negative we can treat the system as though positively charged holes having a positive effective mass were present.

The effective mass can be determined from cyclotron resonance experiments and, at least in principle, we could derive a density of states function from it. However, the experimental values of effective mass apply only to electrons that are near to either a band edge or the maximum energy of electrons in a sample at absolute zero, and can only be used to find the density of states in these regions. We cannot deduce the density of states for the entire band from these measurements. Instead, we must regard the distribution of states in energy as being essentially unknown, but having a distribution near the band edge as given by Eq. (10.4) if the mass is replaced by the effective mass.

We have noted that if we were to plot the points corresponding to permitted values of true momentum in P-space, those points having the same total energy would lie on spherical surfaces. This would not ordinarily be true for crystal momentum. In this case, points of equal energy would form rather general closed surfaces centered around energy minima. These are called *Fermi surfaces*. These energy minima are not necessarily located at the origin, but may be distributed in a manner such as is shown in Fig. 10.3. Solids having energy minima other than at the origin are called *many-valley solids*.

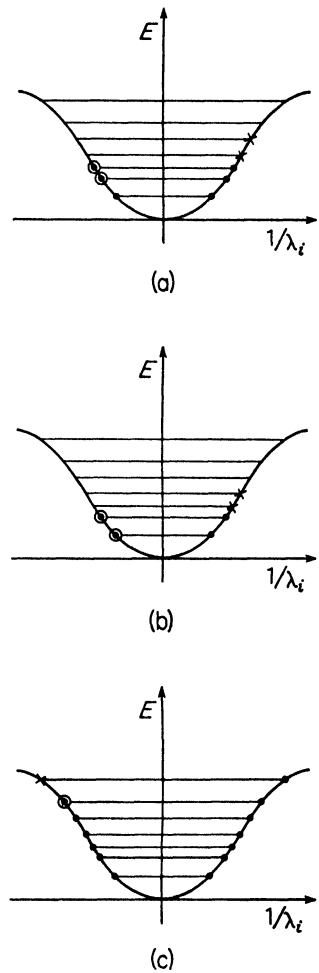


Figure 10.2 Illustration of possible variation of effective mass with energy.

According to Eq. (10.7), the effective mass of a charge carrier having a given energy is equal to the curvature in P-space of the Fermi surface. For this reason, $1/m^*$ will vary with direction in P-space and will be a tensor quantity. This may be seen more readily by writing Eq. (10.7) in the dyad

form

$$\nabla_P \nabla_P E = \frac{1}{\mathbf{m}^*} \tag{10.8}$$

Notice that this dyad is defined in P-space rather than in ordinary space. Although $1/\mathbf{m}^*$ is a tensor in P-space, the effective mass itself is a tensor of rank two in ordinary Euclidean space.

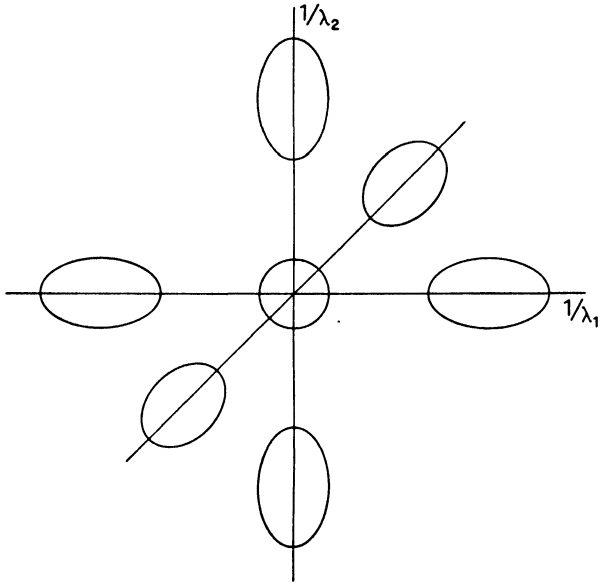


Figure 10.3 Fermi surfaces in a many-valley solid.

To summarize, the elementary zone theory presented in Chap. 5 gives a complete but distorted picture of the distribution of states in energy. We must now regard the actual distribution of states in energy to be beyond any calculation from first principles, but we may find the distribution of states in certain energy ranges experimentally. Charge carriers in a solid may act as though their mass differed from their mass in free space. This behavior is characterized by the material property called the *effective mass* which is normally a tensor quantity.

10.2 The Fermi-Dirac distribution

If the density of states function were known, the next logical step would be to determine how these states are populated by electrons and/or holes. At absolute zero, the electrons of a system will populate the quantum states of lowest energy leaving no vacant states having energies less than the maximum

electron energy. At any temperature above absolute zero, the electrons will participate in the general partition of thermal energy and a certain fraction of these will populate states having energies in excess of that of the highest populated state at absolute zero. Thus the bottom of the band will consist of completely filled states, but in the region near the highest populated state at absolute zero there will be a distribution of partially filled and vacant states. In the paragraphs that follow we shall derive a function enabling us to calculate the probability that a state of prescribed energy will be occupied.

Consider two narrow energy bands, A and B. The rate of transition of electrons from band A to band B is proportional to the number of occupied states in band A and to the number of vacant states in band B. Thus

$$R_{A\to B} = C_1 N(A)\rho(A)N(B)[1 - \rho(B)]$$

Similarly

$$R_{A\to B} = C_2 N(B)\rho(B)N(A)[1 - \rho(A)]$$

where $\rho(A)$ is the probability that a state having an energy A is occupied. At equilibrium

$$R_{A\to B} = R_{B\to A}$$

Hence

$$\frac{\rho(B)}{1 - \rho(B)} = \frac{K\rho(A)}{1 - \rho(A)}$$

The constant K is equivalent to the familiar equilibrium constant of chemical reactions. It would cause too great a digression into kinetic theory to prove it, but the variation of K with temperature is expressed by

$$\frac{d \ln K}{dT} = \frac{B - A}{kT^2}$$

Integration and substitution produces

$$\frac{A}{kT} + \ln \frac{\rho(A)}{1 - \rho(A)} = \frac{B}{kT} + \ln \frac{\rho(B)}{1 - \rho(B)}$$

Since the lefthand side of this equation is a function of only A and T and the righthand side is a function of only B and T, they must both be equal to some function of T which is independent of the energy of either A or B. Thus

$$\ln \frac{\rho(A)}{1 - \rho(A)} = \frac{\zeta - A}{kT}$$

Finally, since A and B are arbitrary energies, we may write

$$\ln \frac{\rho(E)}{1 - \rho(E)} = \frac{\zeta - E}{kT}$$

The reduction to

$$\rho(E) = \exp - \frac{(E - \zeta)}{kT} \tag{10.9}$$

is obvious. $\rho(E)$ is called the *Fermi-Dirac distribution function.* This probability function will be such that the fraction of the electrons having energies between

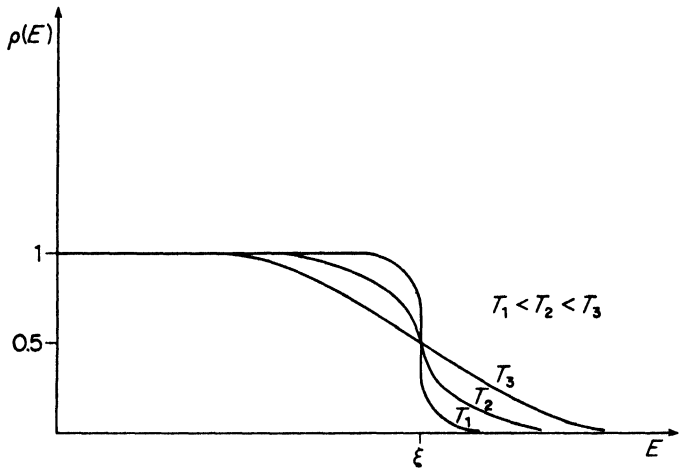


Figure 10.4 The Fermi-Dirac distribution.

E and $E + dE$ will be related to the density of states function through

$$\frac{dn}{n} = \rho(E)N(E)\,dE \tag{10.10}$$

The quantity ζ in Eq. (10.9) is a parameter having the dimensions of energy called the *Fermi energy*, and k is the Boltzmann constant. The physical significance of ζ is that it represents an energy such that the number of occupied states having an energy in excess of ζ is equal to the number of vacant states having an energy less than ζ. In general, the Fermi energy will vary with temperature. The temperature dependence of ζ depends upon the density of states function and the ratio of electrons to states in the system. We shall discuss this point in greater detail later.

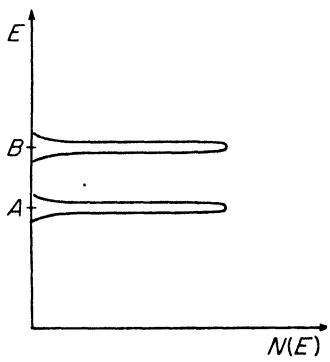


Figure 10.5 Band structure of hypothetical material cited in the adjacent text.

Figure 10.4 shows the variation of $\rho(E)$ with energy for a fixed value of ζ at different temperatures. For a specific application of the Fermi-Dirac distribution function, let us consider a covalent solid whose band structure is shown in Fig. 10.5. This case, of course, is hypothetical and does not correspond to any real solid. Both bands consist of states of nearly the same

energy. For a case such as this, the Fermi level will lie halfway between the bands at a level $(E_A - E_B)/2$. If the bands are due to $_2s$, and $_2p$ states, each band will contain four states per atom. The fraction of the states in the conduction band that are occupied will be

$$\rho(E) = \frac{1}{1 + \exp \dfrac{2E_A - E_A - E_B}{2kT}} = \frac{1}{1 + \exp \dfrac{E_A - E_B}{2kT}}$$

and the concentration of electrons in the conduction band as a function of temperature will be

$$n_e = \frac{4N}{1 + \exp \dfrac{E_A - E_B}{2kT}}$$

where N is the number of atoms per unit volume in the system.

Approximations such as that used in the example cited are fairly common, even for cases where the bands have finite width. The reason for their validity is that most of the vacancies in the valence band will be very close to the top edge of the band and most of the electrons in the conduction band will reside in states close to its bottom edge.

10.3 Electrical conductivity

From a purely phenomenological point of view, the electric current density at a point in a medium free of temperature gradients is related to the potential gradient at the point by

$$\mathbf{J}_e = \mathbf{C}_{ev} \cdot \nabla V \tag{10.11}$$

without regard for the mechanisms of conduction. The magnitude of the components of the electrical conductivity tensor, $\mathbf{C}_{ev}$, is determined by two different factors: the concentration of charge carriers present, and their mobility. Assuming both positive and negative charge carriers are present, the conduction is expressed by

$$\mathbf{C}_{ev} = e(n^-\mu^- + n^+\mu^+) \tag{10.12}$$

where e refers to the charge on the carrier, the n's to the concentration of carriers, and the μ's to their mobility.

The concentration of charge carriers depends upon entirely different factors in metals than it does in semiconductors, but the mobilities are determined by the same factors in both cases. Therefore, we shall devote the next section exclusively to discussion of mobility; then we shall consider metals and semiconductors separately when discussing the factors that influence concentration.

10.4 The mobility of charge carriers

In order to change direction, the electrons in a lattice must interact either with each other or with the lattice. These interactions are governed by the usual conservation of momentum and energy laws and would be rather rare in a perfect lattice. However, even if a lattice is otherwise perfect, phonons will be present at nonzero temperatures in sufficient concentrations to provide a mechanism for the charge carriers to interact and equilibrate with the lattice. Thus at any given temperature the distribution of electrons in energy will be constant, but the individual electrons will undergo frequent changes in momenta and energy as they collide with phonons, dislocations, impurities, and other sources of lattice strain. During the intervals between such events, if an electric field is present they will experience an acceleration because of the force exerted upon them by the field. In terms of the change of crystal momentum, this acceleration may be expressed by the equation

$$\frac{d\mathbf{P}}{dt} = \nabla V \cdot \frac{1}{\mathbf{m}^*} \tag{10.13}$$

Over a period, long in comparison with the average time between collisions, this will impose upon the electron a net drift velocity in a direction determined by the direction of the electric field and the $1/\mathbf{m}^*$ tensor. The mobility

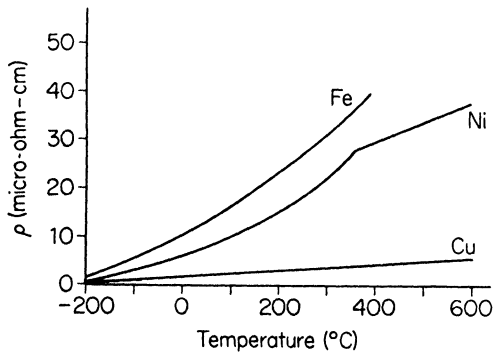


Figure 10.6 Electrical resistivity of some single crystal cubic metals. (Domenicali and others, *J. App. Phys.* **26,** 377 (1955).)

is defined as the value of this drift velocity produced by a unit potential gradient. We note that it will be directly proportional to the average time between collisions, and will be inversely proportional to the effective mass. If we make the very reasonable assumption that the scattering process is truly random, then the mobility will be a tensor of the same rank as $1/\mathbf{m}^*$ and will have components that are proportional to those of $1/\mathbf{m}^*$ with the same proportionality factor applying to all of the components in any coor-

dinate system. However, this proportionality constant is a function of the concentration of scattering centers. Thus the mobility will be decreased by an increase in the density of dislocations, by increasing concentrations of impurity atoms, and by increasing temperature. Figure 10.6 shows the variation with temperature of the electrical conductivity of some single crystal cubic metals illustrating the effect of thermal scattering.

10.5 Metals

The fundamental criterion of a metal is that those quantum states which are populated by electrons at absolute zero must be immediately contiguous to vacant states. This can be achieved in different ways depending upon the band structure of the material under consideration and the ways the bands are populated by electrons. We recall that the bands consist of quantum states distributed in very small energy increments over a range of energies. However, each state is derived from a particular combination of permitted values of crystal momentum. Each of the bands of a solid consists of all of those states corresponding to permitted values of momentum that are enclosed between the boundaries of Brillouin zones. Thus if we plot permitted values of crystal momentum in *P*-space, those points inside the first zone will correspond to the quantum states in the lowest band. Those points representing allowed values of crystal momentum which lie between the boundaries of the first and second Brillouin zones correspond to the energy states in the next higher band, etc. The core electrons of an atom cannot form continuous bands, and we are speaking here only of those electrons in the highest energy levels of the individual atoms. When a large number of atoms are condensed to form a crystalline solid, and the temperature is reduced to absolute zero, the electrons completely fill the states of lowest energy in the solid until all are accommodated. These states have a twofold degeneracy, therefore their momenta will paired and opposed. Consequently, the net transport of charge is zero. In order to carry current, the sum of the momenta of all of the electrons near the Fermi level must be distributed over a range of partially filled states, and only these can be considered as conduction electrons. However, it is not necessary that this distribution be accomplished by temperature. Even very small electric fields are capable of moving an electron from one state to the next higher state, provided that it is not already occupied by an electron. If the valence band of a metal were only half filled with electrons, then, in principle, all of the electrons in the band could become conduction electrons. In practice, the high fields necessary to achieve this could not be realized in a metal, and only those electrons having energies near the Fermi level actually participate in conduction. The concentration of these electrons depends upon the field intensity.

Those metals which are in group I of the periodic table only have half as

many electrons as there are states in the band corresponding to the first Brillouin zone, and those in group III fill all of the states corresponding to permitted crystal momenta inside the first zone, and half of those in the second zone. In both of these cases there are empty states which electrons can move into after receiving only very small increments of energy from an electric field. When a field is present, then, electrons may be accelerated, and as their kinetic energies increase they will move to progressively higher states in the band. Normally, they will be scattered by phonons, dislocations, etc. long before they achieve energies near the top of the band. Thus it would be very improbable that an electron would have a momentum lying near the surface of the zone.

With group II and transition metals the situation is more complex. The highest principal quantum level of the group II metals has two electrons. The number of states corresponding to permitted values of crystal momentum enclosed by the first Brillouin zone is just twice the number of atoms in the system. Thus the lowest band would have exactly the number of states necessary to accommodate all the electrons. In a completely filled band the electrons cannot conduct because of their paired momenta. In the case of the group II and transition metals some of the energy states in the second Brillouin zone have a lower energy than the highest energy levels in the first zone. Consequently, some of the electrons will reside in the second zone and there will be some vacancies in the first zone. This overlap of the zones provides the necessary empty states near the Fermi level which are responsible for the properties typical of a metal.

The band structure and the population of the energy states for such a metal are shown in Fig. 10.7. The electrons near the Fermi level will all have crystal momenta near a zone boundary. Those electrons in states corresponding to crystal momenta belonging to the second Brillouin zone can be accelerated by an electric field in a normal manner, but those in states belonging to the first zone cannot normally pass into states belonging to the second zone without changing the direction and magnitude of their crystal momentum. In order to do so, they would have to interact with a lattice defect and would lose energy in the collision. For this reason, only a few electrons are able to pass into the upper band.

Thus the electrons in the first Brillouin zone are fairly well confined to energy states below the top of the valence band. In this region, the electrons will have negative effective masses and will be accelerated by the field in directions opposite to those having positive effective masses. Consequently, there will also be a certain amount of hole conduction in these metals. In either case, the concentration of charge carriers will be low and the conductivity will be correspondingly low.

We have noted that it is only the electrons near a band edge that participate in conduction. Usually these will be distributed over such a narrow

range of energies that they all have essentially the same effective mass. In the case of the group II and transition metals, conduction occurs through the action of electrons in two different bands, thus two different effective masses should be distinguishable. Indeed, conducting electrons in the valence band will have a negative effective mass which is manifested experimentally as though a hole having a positive effective mass were present.

In summary, we have explained qualitatively the observed lower conductivity of group II and transition metals as compared with other metals.

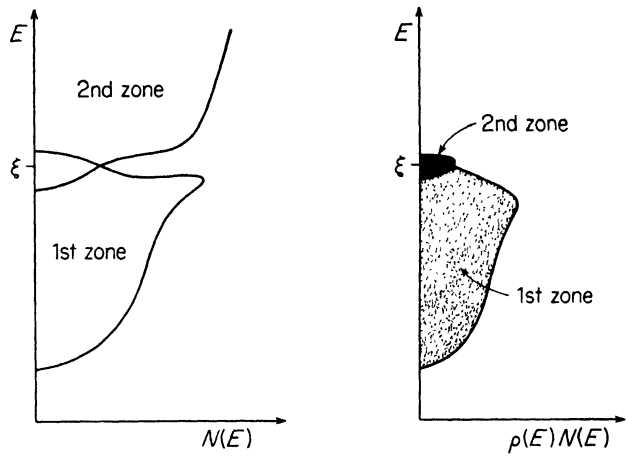


Figure 10.7 Density of states functions when there are overlapping bands (left), and population of the bands by electrons at some non-zero temperature (right).

We have also indicated the nature of some problems which presently command the interest of solid-state researchers. The concentration of charge carriers in metals is not a well-defined quantity and can be obtained formally only for cases sufficiently simple to allow the mobility to be measured independently.

10.6 Semiconductors

In Chap. 5, we established criteria for distinguishing metals from semiconductors in terms of their band structure. To recapitulate, a semiconductor is a solid which, at absolute zero, has a band completely filled with electrons. Above this band but separate from it is another band which is empty at absolute zero. Between the bands there is a gap in energy which contains no states that can be occupied by electrons. This is shown schematically in Fig. 10.8. The conduction band is composed of the antibonding states discussed in Chap. 5.

Even at very low temperatures, there will be some phonons present in the lattice which are sufficiently energetic to partially rupture a bond between atoms. As the temperature increases, their concentration increases until at any nonzero temperature a dynamic equilibrium will be attained between partially ruptured bonds freeing electrons and partially ruptured bonds reacting with free electrons to heal themselves. In terms of the density of states and the Fermi-Dirac distribution functions, the population of electrons in the upper band and the population of vacancies in the lower band might be shown schematically as in Fig. 10.8. If the lower part of the conduction band has the same density of states function as the upper part of the valence band, the Fermi energy would not vary with temperature. Assuming this is

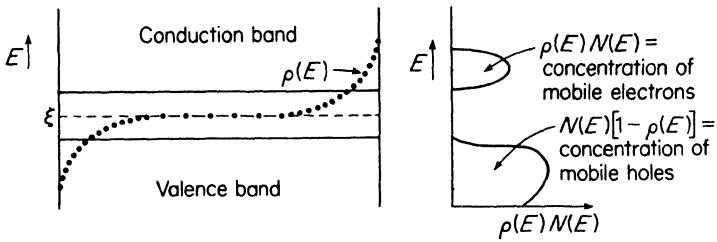


Figure 10.8 Band structure of a semiconductor showing (by the dashed line) the Fermi level and (by the dotted line) the Fermi-Dirac distribution function. The adjacent figure is the population density of electrons and holes.

true and making certain of other assumptions about the average kinetic energy of the electrons in the conduction band gives the following relationship between the concentration of electrons in the conduction band, the temperature, and the gap between the bands

$$n_{e^-} = 2\left(\frac{2\pi m^* kT}{h^2}\right)^{3/2} e^{-E_g/2kT} \tag{10.14}$$

When electrons are excited from the valence band to the conduction band, they leave holes in the valence band. In the case treated in the previous paragraph, there is one such hole in the valence band for each electron in the conduction band. Such a semiconductor is called an *intrinsic semiconductor*.

Now, on the basis of the ideas developed in this section, let us try to predict the way that a component of the electrical conductivity tensor will vary with temperature. From Eq. (10.14) we may approximate the concentration of electrons in the conduction band, and since this is equal to the concentration of holes

$$C_{ev} = 4e\left(\frac{2m^* kT}{h^2}\right)^{3/2} (\mu^- + \mu^+) e^{-E_g/2kT} \tag{10.15}$$[1]

[1] Ordinarily, the effective mass of electrons is different from that of holes. In this case, we have assumed that the density of states function near the lower edge of the conduction

where μ is the mobility of the charge carriers. Taking the logarithm of both sides and differentiating gives

$$\frac{d \ln C_{ev}}{d(1/T)} = \frac{3}{2} \frac{d \ln}{d(1/T)} \left(\frac{2\pi m^* kT}{h^2} \right) - \frac{E_g}{2k} + \frac{d \ln}{d(1/T)} (\mu^- + \mu^+)$$

Since the mobilities are about linear with $1/T$, the first and last terms will be small in comparison with the middle term and may be neglected. Consequently, we would expect that a plot of electrical conductivity versus $1/T$ would be approximately linear and would have a slope equal to $(E_v - E_c)/2k$. This is typical of the behavior of the conductivity of an intrinsic semiconductor and provides a reasonable estimate of the gap between the conduction and valence bands.

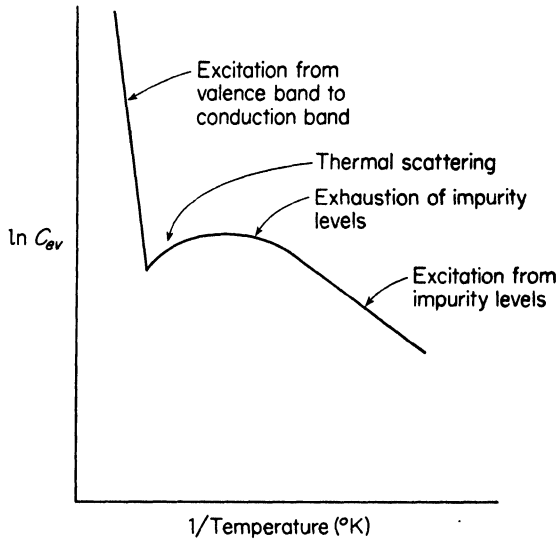


Figure 10.9 Variation of electrical conductivity with temperature for a hypothetical impurity semiconductor.

Strictly speaking, the foregoing arguments are oversimplified, since it was assumed that there were relatively few charge carriers and that the band structure could be represented by a single energy surface. Many of the compound semiconductors such as InSb, GaAs, PbTe, Ag_2Se, and the like have such small energy gaps that they become degenerate at very low temperatures. This means that the density of charge carriers is sufficiently high that an appreciable proportion of these will reside in states in pairs having opposed momenta. These will act very much like metals in their variation of resistivity with temperature. Figure 10.9 shows the variation of conduc-

band is the same as that near the upper edge of the valence band. This requires that the effective masses of the electrons and holes be identical.

tivity with temperature for a typical semiconductor; Table 10.1 lists energy gaps and mobilities for a variety of materials.

Table 10.1 *Energy gaps and mobilities of certain materials**

Material	E_g	*Mobility* (cm^2/v-*sec*)	
Diamond	6.0	1,200	1,800
Si	1.10	400	1,600
Ge	0.7	1,800	3,800
InSb	0.18	1,250	77,000
PbTe	0.30	..	17,000

* From Charles Kittel, *Introduction to Solid State Physics* (2nd ed.; New York: John Wiley & Sons, 1956).

P- and n-type semiconductors. Figure 10.10 is a schematic diagram of the atoms in the germanium lattice, where the atoms are represented as lying in a plane. Each atom is bonded to its four nearest neighbors by covalent bonds, each bond containing two electrons, one contributed by each atom participating in the bond. In the center of the sketch is an unlabeled atom. Let us assume that this atom is a group V atom containing five electrons in its outer principal quantum level. Four of these will participate in bonds with germanium atoms, but the fifth will not, remaining instead rather loosely bonded to the impurity atoms. Sufficiently energetic phonons can excite electrons from the valence band to the conduction band, but phonons of much lower energy are capable of freeing the fifth electron from the impurity atom. If such a phonon should collide with this electron, it would enter the conduction band and be free to move through the lattice leaving a positively charged immobile impurity ion behind. By this mechanism it is possible to have conduction electrons in a semiconductor without having an appreciable number of holes present. If the impurity concentration is about one part in 10^5, we may think of them as contributing a series of isolated energy states that lie in the forbidden region near the conduction band as shown by the dotted line in Fig. 10.11.

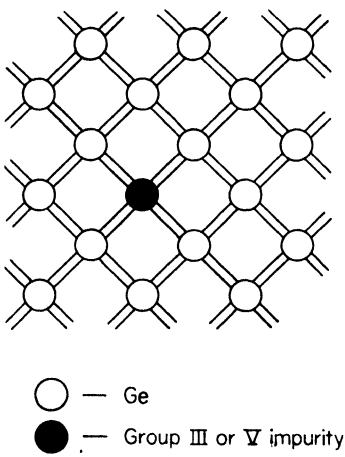


Figure 10.10 Schematic lattice.

At a temperature where the excitation of electrons out of the valance band is negligible, the Fermi level will lie between the impurity states and the bottom of the conduction band. Again we may make the approximation that the electrons excited from the impurity levels will be distributed in states very near the bottom of the conduction band, and that the Fermi level will lie halfway between the conduction band and the impurity levels and

will be independent of temperature. Then the concentration of electrons will be approximately

$$n_e = N_D \left(\frac{2\pi m^* kT}{h^2}\right)^{3/2} e^{-[(E_c - E_D)]/2kT} \tag{10.16}$$

Again we may predict qualitatively how the conductivity might vary with temperature. At very low temperatures, the conductivity will be due to excitation of electrons from the impurity levels. The slope of the plot of $ln\ C_{ev}$ versus $1/T$ will approximate the energy of the gap between the impurity level and the conduction band. Eventually, as the temperature is raised, most or all of the impurities will have ionized, and increasing temperature will not increase the concentration of electrons. Instead, the mobility decrease accompanying the increase in temperature will begin to dominate the scene, and the conductivity will start to decrease with further temperature increase.

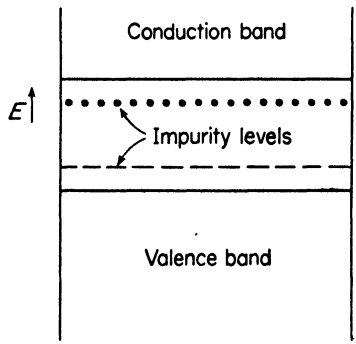


Figure 10.11 Band structure of a semiconductor indicating the location of the donor and acceptor levels in the forbidden region.

Finally, excitation from the valence band will cause a rapid increase in conductivity with temperature. These various stages are shown in Fig. 10.9. Since the conduction is due to electrons, this type of semiconductor is called an *n-type semiconductor*.

Now let us return to Fig. 10.10 and consider the situation that would ensue if the unlabeled atom in the matrix were a group III impurity rather than a group V. In this case, only three complete bonds can be formed between the impurity atom and its germanium neighbors. The fourth bond is incomplete and consists of only one electron. This incomplete bond will be able to capture electrons excited from the valence band. This is accomplished by collision with a phonon having insufficient energy to excite the electron into the conduction band. In this case, the impurities act like electron acceptors having energies lying in the forbidden region near the valence band as shown by the dashed line in Fig. 10.11. When this occurs, the acceptor atom becomes an immobile negative ion and results in a mobile

hole in the valence band. This type of semiconductor is called a *p-type*, because the conductivity is due to positively charged holes. The variation of conductivity with temperature is similar to that of an n-type semiconductor.

10.7 Applications

Possibly the most famous devices whose operation is explainable in terms of the ideas developed in the preceding sections are transistors and solid-state diodes. Both of these are normally made from single crystals of a semiconductor such as silicon or germanium. In the case of the diode, different types of impurities are allowed to diffuse into the crystal from opposite directions. Under properly controlled conditions, the concentration gradient at the front of the diffusing material can be kept very abrupt, and when the impurities meet inside the crystal the process is stopped. In this way, part of the

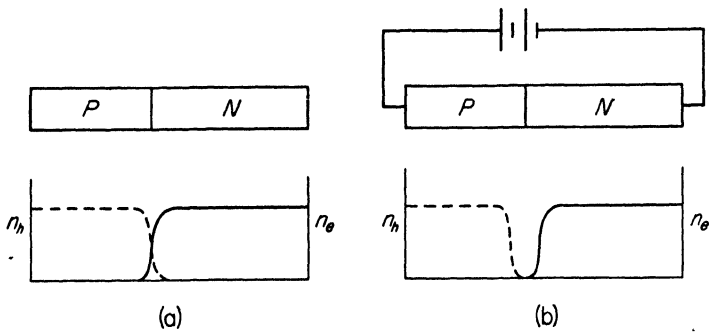


Figure 10.12 Distribution of electrons and holes across a p-n junction with and without a bias voltage.

crystal can be rendered n-type and part p-type with an abrupt transition from one to the other at an interface extending along a plane passing through the crystal. Metallurgically both sides of the crystal are identical,[1] and the concentration of the impurities is so low that it would even be difficult to detect the slight difference in chemical composition of the two sides by spectroscopic means. Electrically, though, they differ by virtue of the fact that the majority of the charge carriers on one side of the boundary are electrons while the majority on the other are holes.

Whenever a concentration gradient exists, the potentiality of diffusion also exists. Thus electrons will be driven to the p-side of the crystal by their concentration gradient, and holes similarly will diffuse into the n-side. These diffusion processes will finally reach an equilibrium condition in which

[1] This statement needs some qualification. There will usually be different chemical reactivities; consequently, an etch will develop a step at a junction, which casts a shadow and makes microscopic examination of the junction possible.

the p-side will have a net negative charge and the n-side will have a net positive charge. The resulting electric field across the boundary will oppose further diffusion. At this point, the concentrations of holes and electrons inside the crystal will follow the curves shown in Fig. 10.12a. Now if the terminals of the diode are connected with the n-side to the positive terminal of a battery and the p-side to the negative terminals, the concentration curves of both electrons and holes will be shifted in the manner shown in Fig. 10.12b, leaving only a very small concentration of charge carriers of either side in the region of the junction. Thus only a very small current will be able to flow.

On the other hand, if the battery connections are reversed, there will be no region in the crystal depleted of the charge carriers and current will flow. This, in essence, is the way the solid-state diode operates. Incidently, when current does flow in the forward direction, electrons and holes will annihilate each other near the junction with the liberation of heat. This is

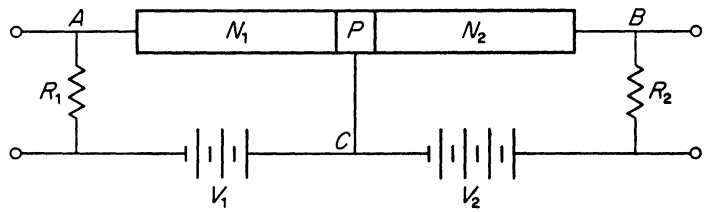


Figure 10.13 A possible circuit for using an n-p-n transistor as a power amplifier.

not an ohmic heating, but is a thermoelectric effect which we shall treat in the next chapter.

In its simplest form, the transistor consists of a single crystal containing either two n-type regions separated by a very thin p-type region, or of two p-type regions separated by a very thin n-type region. Normally when an electron and a hole collide the result is mutual annihilation. Consequently, holes in an n-type region and electrons in a p-type region have only a very limited life. The central region in either a p-n-p or an n-p-n transistor must be sufficiently thin so that minority carriers can diffuse through it with only a very small probability of being annihilated.

We examine the mechanics of the transistor action by considering an n-p-n type in a circuit such as that shown in Fig. 10.13. The battery V_1 will cause electrons to move from the n_1-region toward the p-region. Because of its thinness they will pass straight through it with only a relatively small number being annihilated by holes or passing down the connection from the p-region to the junction between batteries. Once they enter the n_2-region, they are moved through the region by the field produced by the battery V_2. The battery V_2 by itself cannot produce any current flow since the p-n_2

junction blocks any current flow that the potential difference between points A and B might have been expected to pass. Thus i_c is less than but almost equal to i_e and the latter is controlled by the potential difference between points A and C. We are choosing current flow to be in the direction of electron flow, in direct contradiction to a well-established but archaic convention.

The ultimate limit upon the operation of the device is that the voltage drop across R_1 must be less then V_1 and the voltage drop across R_2 must be less than V_2. If we choose to make V_2 large in comparison with V_1, then the ratio of the output voltage to the input voltage will be only slightly less than V_2/V_1 and the power dissipated in R_2 will also be correspondingly greater than the power dissipated in R_1. Furthermore, the output is directly controlled by the voltage across R_1. Consequently, we could apply an a-c signal across the input terminals whose peak-to-peak value would be limited to values less than V_1, and we would have a replica of this signal appearing across R_2 except that its magnitude would be larger and its capability of delivering power would also be larger. Thus the circuit of Fig. 10.13 is capable of acting as a voltage and power amplifier. Although this circuit is capable of operating, it will not operate well and no sane circuit designer would ever use it. The transistor has many advantages over its vacuum tube counterpart, but one significant disadvantage is that the output load is reflected back to affect the input impedance to a much greater degree than it would in an equivalent vacuum tube circuit. It is this factor that often makes transistor circuits more elaborate than one expects from a direct consideration of the mechanics of transistor action.

PROBLEMS

10.1 If we accept the equation

$$g(E)dE = \frac{\pi}{2}\left(\frac{8m^*}{h^2}\right)^{3/2}(E - E_g)^{1/2}dE$$

as a proper expression of the density of states function in a semiconductor, derive the equation

$$n^- = 2(2\pi m^* kT)^{3/2} \exp(\zeta - E_g)/kT$$

10.2 Assuming that $E_g \gg (E_g - E_D)$ and that all of the conduction electrons can be considered to be residing in states very near to the bottom of the conduction band, where would the Fermi level lie when 50 per cent of the donors in an n-type semiconductor were ionized?

10.3 Given the following data for a p-type semiconductor, calculate the hole concentration at 100°K.:

$N_A = 10^{15}$ atoms/cm^2
$m^* = 0.9$ m
$E_A = 0.05$ electron volts

11

CROSS CONDUCTIVITIES

11.1 Conductivities and resistivities

When the various classes of material properties were discussed in Chap. 6, it was pointed out that whenever a material sustains a gradient of a scalar parameter the potentiality exists for this gradient to cause a primary flow of the corresponding extensive parameter and possibly to cause secondary flows of other extensive quantities. The flow of these quantities can be characterized by a set of current density vectors and the gradients can be characterized by either of two sets of rank two tensor quantities called *resistivities* and *conductivities*. Choosing electricity and entropy as being the most important examples in which such cross effects are observed, we can define the appropriate resistivities and conductivities by the equations

$$-{}_{e}\mathbf{J} = {}_{e}\mathbf{C} \cdot \nabla V + {}_{eT}\mathbf{C} \cdot \nabla T \tag{11.1a}$$

$$-{}_{s}\mathbf{J} = {}_{eT}\mathbf{C} \cdot \nabla V + {}_{T}\mathbf{C} \cdot \nabla T \tag{11.1b}$$

and

$$-\nabla V = {}_{e}\boldsymbol{\varrho} \cdot {}_{e}\mathbf{J} + {}_{eT}\boldsymbol{\varrho} \cdot {}_{s}\mathbf{J} \tag{11.2a}$$

$$-\nabla T = {}_{eT}\boldsymbol{\varrho} \cdot {}_{e}\mathbf{J} + {}_{s}\boldsymbol{\varrho} \cdot {}_{s}\mathbf{J} \tag{11.2b}$$

In the above equations, the $\boldsymbol{\varrho}$'s are the resistivities, the $\mathbf{C}$'s conductivities, and the subscripts preceding the symbols designate these as thermal (T), electrical (e), or thermoelectric (eT). Since either pair of equations would suffice to relate the gradients and the current densities, the conductivities and resistivities are not independent. Using the same techniques developed in Chap. 6 to relate compliances and rigidities, it can be shown that the following set of equations serve to relate the resistivities and conductivities to one another:

$$
\begin{aligned}
{}_{e}\mathbf{C} &= \frac{{}_{T}\boldsymbol{\varrho}}{{}_{e}\boldsymbol{\varrho} : {}_{T}\boldsymbol{\varrho} - {}_{eT}\boldsymbol{\varrho} : \boldsymbol{\varrho}_{eT}} && \text{(a)} \\
{}_{eT}\mathbf{C} &= \frac{-{}_{eT}\boldsymbol{\varrho}}{{}_{e}\boldsymbol{\varrho} : {}_{T}\boldsymbol{\varrho} - {}_{eT}\boldsymbol{\varrho} : {}_{eT}\boldsymbol{\varrho}} && \text{(b)} \\
{}_{T}\mathbf{C} &= \frac{{}_{e}\boldsymbol{\varrho}}{{}_{e}\boldsymbol{\varrho} : {}_{T}\boldsymbol{\varrho} - {}_{eT}\boldsymbol{\varrho} : {}_{eT}\boldsymbol{\varrho}} && \text{(c)} \\
{}_{T}\boldsymbol{\varrho} &= \frac{{}_{e}\mathbf{C}}{{}_{e}\mathbf{C} : {}_{T}\mathbf{C} - {}_{eT}\mathbf{C} : {}_{eT}\mathbf{C}} && \text{(d)} \\
{}_{eT}\boldsymbol{\varrho} &= \frac{-{}_{eT}\mathbf{C}}{{}_{e}\mathbf{C} : {}_{T}\mathbf{C} - {}_{eT}\mathbf{C} : {}_{eT}\mathbf{C}} && \text{(e)} \\
{}_{e}\boldsymbol{\varrho} &= \frac{{}_{T}\mathbf{C}}{{}_{e}\mathbf{C} : {}_{T}\mathbf{C} - {}_{eT}\mathbf{C} : {}_{eT}\mathbf{C}} && \text{(f)}
\end{aligned}
\qquad (11.3)
$$

The fact that the conductivities and resistivities are tensor quantities is in itself sufficient to cause some rather unexpected phenomena. Even if the cross conductivities were zero, in which case there would be no interaction between thermal and electric phenomena within a given material*, we would still have the gradients interacting with the material in such a way that they would not necessarily coincide in direction with the current densities they produce. In anisotropic materials, the gradient and the resultant current density coincides in direction only in the special case where the gradient lies in the direction of a principal axis of the conductivity tensor. This fact can be illustrated by considering a material in which the thermoelectric cross conductivity is absent. Then Eq. (11.2a) becomes

$$-\nabla V = {}_{e}\boldsymbol{\varrho} \cdot {}_{e}\mathbf{J} \qquad (11.4)$$

We are free to choose a coordinate system with one axis in the direction of the current density vector, and in the general case the righthand side of the equation will involve three components of the gradient. The exception is the case in which ${}_{e}\rho_{ij} = 0$ for $i \neq j$. Then the direction of the current density, the gradient, and one principal axis of the conductivity tensor coincide. In this connection, Neumann's principle is capable of providing some information about the orientation of the principal axes of the tensors with respect to the crystallographic axes. All rank two tensor properties of orthorhombic materials will have their principal axes coincident with the crystallographic axes. For crystals having lower symmetry this rather sweeping statement does not apply, but even so, it provides a powerful technique for interpreting experimental data obtained from systems having sufficient symmetry.

* There could still be interactions arising from processes occurring at the junctions of two different materials.

When the thermoelectric cross effects are not negligible, the problem of evaluating the components of the conductivity or resistivity tensor experimentally is enormously complicated by the difficulty in specifying the boundary conditions of the sample during the experiment. Under these circumstances the passage of an electric current will produce temperature gradients, and the passage of an entropy current will produce potential gradients. If we attempt to hold either the electric current or the entropy current to zero by isolation of the sample, then we must look seriously at the question of how effective the isolation is if instrument probes are brought into physical contact with the sample. Difficulties such as these can be ameliorated to some degree by a judicious use of sample geometry.[1] In particular, the use of a planar sample will at least assure that one component of the electrical current is zero, and by using long thin specimens two components of the current can be limited to essentially zero. Figure 11.1 shows a particularly useful type of sample configuration for making measurements of this type.

For purposes of illustration let us consider how measurements of the components of electrical conductivity could be made upon such a sample. The experimental conditions are as follows:

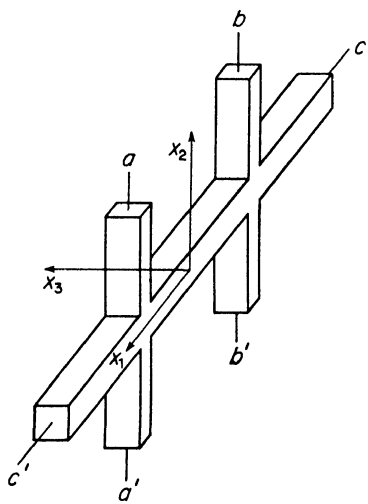


Figure 11.1 Single crystal cut for resistivity measurements.

1. The sample is a single crystal having orthorhombic symmetry. The *a* crystallographic axis is oriented parallel to the x_1 axis shown in the figure and the *c* axis is coincident with the x_3 axis.

2. The entire sample is immersed in a thermostat and the component of temperature gradient in the x_1 direction is assumed to be zero.[2]

3. A measured electric current is supplied through terminals cc'.

4. Potential measurements between aa', bb', and ab can be made with an instrument having essentially an infinite input impedance.

The potential gradient between the cross arms may be considered constant, and the electrical current density in this region may be considered to be equal to the total current divided by the cross-sectional area.

If C_{11}, C_{22}, and C_{33} are the principal components of the electrical con-

[1] Actually, the cross conductivity is usually so low that this is rarely much of a problem in experimental procedure.

[2] Since the current will transport heat, this may be hard to achieve.

ductivity tensor, the matrix of the components of this tensor in the experimental coordinate system is

$$\begin{bmatrix} C_{11} & 0 & 0 \\ 0 & C_{22} & 0 \\ 0 & 0 & C_{33} \end{bmatrix}$$

We have omitted the subscript e with the understanding that C_{ij} is a component of the electrical conductivity. The electric current density is limited by sample geometry to the x_1 direction and is easily measured. The x_1 component of the temperature gradient is held at zero, and the radial temperature gradient is symmetrical about the center of the sample, hence its effect is self-canceling. Consequently Eq. (11.2a) reduces to

$$J_1 = -C_{11}\nabla_1 V.$$

$\nabla_1 V$ is deduced from the potential difference between terminals a and b, thus permitting the evaluation of C_{11}. In this case, no potential difference appears between terminals aa' or bb'. Had the crystallographic axes been given some other orientation, a transverse potential would have been developed.

Now let us use the same experimental geometry, but change the experimental conditions to illustrate a case where a cross effect may be manifested. The new experimental conditions are

1. The sample is a single crystal having orthorhombic symmetry, and the a axis is oriented 45° above the x_1 axis. The b axis coincides with the x_3 axis of Fig. 11.1.
2. The electric current density is zero, and a measured entropy current is flowing through the sample in the x_1 direction.
3. The temperature gradient is considered to be constant in the x_1 direction and zero in the other directions. The magnitude of the temperature gradient is determined by temperature measurements at the cross arms. The electric potential difference between points aa' and ab is measured.

The components of the thermoelectric resistivity in a principal coordinate system are ρ_{11}, ρ_{22}, and ρ_{33}. (Again we have omitted the subscripts eT to simplify notation.) In the experimental coordinate system, the matrix of the components of this tensor is

$$\begin{bmatrix} \frac{\rho_{11}+\rho_{22}}{2}+\rho_{33} & -\frac{\rho_{11}+\rho_{22}}{2} & 0 \\ -\frac{\rho_{11}+\rho_{22}}{2} & \frac{\rho_{11}+\rho_{22}}{2}+\rho_{33} & 0 \\ 0 & 0 & \rho_{33} \end{bmatrix}$$

Since ${}_s\mathbf{J} = 0$, and the only nonzero component of ${}_s\mathbf{J}$ is ${}_sJ_1$, Eq. (11.1a) can

be expanded to

$$-\nabla_1 V = \left(\frac{\rho_{11} + \rho_{22}}{2} + \rho_{33}\right) {}_s J_1$$

$$-\nabla_2 V = -\frac{\rho_{11} + \rho_{22}}{2} {}_s J_1$$

$$-\nabla_3 V = 0$$

The gradient components, $-\nabla_1 V$ and $-\nabla_2 V$, are deduced from potential measurements between aa' and ab; since ${}_s\mathbf{J}$ is known, ρ_{33} and $(\rho_{11} + \rho_{22})/2$ can be evaluated. By making a series of such experiments using different orientations of the crystal, all of the cross-resistivity components can be evaluated.

11.2 Thermoelectricity

Whereas the cross conductivities and resistivities discussed in the previous section are properly considered to be thermoelectric effects, normal usage of the term *thermoelectricity* refers to a more complex phenomenon which is manifested by two different materials joined to form a thermoelectric circuit. The cross conductivities play some part in this phenomenon, but potentials developed at the interfaces between the materials play the dominant role.

When two different materials are joined as shown in Fig. 11.2 and the junctions between them are held at different temperatures, a voltage will be developed across the terminals A and B. If a direct electrical connection is made between the terminals, an electric current will flow and will abet the flow of heat from the hot junction to the cold junction. On the other hand, if a battery is inserted between A and B with the polarity such that it causes a reversal of this electric current, the current will oppose the net flow of heat from the hot junction to the cold junction. If sufficiently large, this will cause a net transfer of heat from the cold junction to the hot junction. In addition to the heat absorbed or liberated at the junctions, there will be an absorption or liberation of heat in those parts of the circuit arms within the temperature gradient. This is linked to the flow of electricity around the circuit.

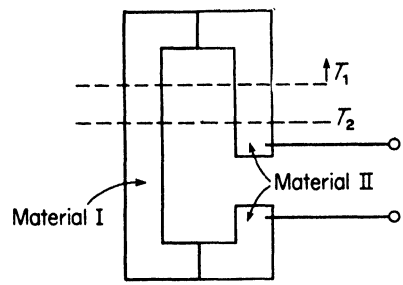


Figure 11.2 A thermoelectric circuit.

In characterizing these thermoelectric phenomena three coefficients are commonly distinguished. Two of these may be conveniently defined in terms of the circuit shown in Fig. 11.2 with the proviso that no part of the temperature gradient lies between the terminals or includes the terminal wires. We shall discuss these two first.

The Seebeck coefficient. When the junctions of Fig. 11.2 are held at different temperatures, the Seebeck coefficient α is defined by the equation

$$\alpha = \operatorname{Lim}_{\Delta T \to 0} \frac{\Delta V}{\Delta T} \tag{11.5}$$

where ΔV is the terminal voltage and $\Delta T = T_1 - T_2$.

The coefficient α is temperature dependent and is a composite property of *both* materials. Since the electric fields set up by the temperature gradient act in the same direction along each leg of the loop, they oppose one another, and the Seebeck coefficient of the pair of materials will be equal to the difference between the absolute Seebeck coefficients of the individual materials

$$\alpha = \alpha_1 - \alpha_2$$

For finite temperature differences between the junctions, the terminal voltage and the Seebeck coefficient are related through

$$V = \int_{T_1}^{T_2} \alpha \, dT$$

We explain later in this section how absolute Seebeck coefficients of the individual materials may be evaluated.

The Peltier coefficient. If a variable capacitor were connected between the terminals in Fig. 11.2, we could cause an electric charge of arbitrarily low value to flow in either direction by varying the capacitance. If the terminal voltage is V, then, when a charge dq is passed around the loop, a quantity of heat dQ will be absorbed at one junction and discharged at the other, depending upon the direction of the flow of charge. In varying the capacitor, a quantity of mechanical work equal to $V\,dq$ will have to be either done upon the capacitor control or absorbed reversibly from the capacitor. If work is done on the capacitor, the heat will be transferred from the cold junction to the hot junction. If heat were transferred from the hot junction to the cold one, the accompanying flow of charge would accumulate on the capacitor plates, which would experience additional electrostatic forces in such a direction that the capacitor could do mechanical work as the plates were readjusted to keep the voltage constant. This transfer of heat is a reversible process superimposed upon an irreversible flow of heat along the branches of the loop not linked to the flow of electricity. Since it is reversible, we may invoke the second law of thermodynamics and write

$$\Delta V \, dq = dQ \left(\frac{T_1 - T_2}{T_1} \right) = dQ \frac{\Delta T}{T_1} \tag{11.6}$$

The *Peltier coefficient* is defined by the equation

$$\pi \, dq = dQ \tag{11.7}$$

for the limiting case where $T_2 \longrightarrow T_1$. Substituting (11.7) into (11.6) gives

$$\frac{\pi}{T} = \text{Lim}\,\frac{\Delta V}{\Delta T} = \alpha \tag{11.8}$$

Again, π is a temperature-dependent composite property of both materials, and is related to the absolute Peltier coefficients of each material by

$$\pi = \pi_1 - \pi_2 \tag{11.9}$$

The Thompson coefficient. If a part of any material contains a temperature gradient, the passage of an electric charge through the material will cause a reversible liberation or absorption of heat in the region lying within the gradient. This is not to be confused with the irreversible Joule heating caused by an electric current, and it is distinguished from the Peltier heat which is concerned only with the reversible absorption or liberation of heat at the *junctions*. The *Thompson coefficient* is defined as the reversible heat liberated per unit length of material by the passage of a unit charge through a unit temperature gradient. We may define the term more precisely by considering the situation shown in Fig. 11.3. This shows a portion of a solid, part of which is in a temperature gradient but which is uniform in temperature above and below the dashed lines. The curve C is an arbitrary curve lying inside the solid. When a charge dq passes through the material, the heat liberated in the region containing the gradient is related to the Thompson coefficient by the line integral

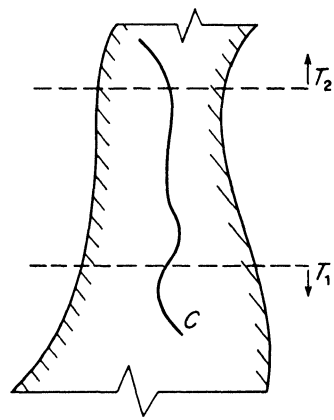


Figure 11.3 A portion of a thermoelectric circuit showing an arbitrary curve C extending from a region of temperature T_1 to a region T_2 inside a material. The line integral of Eq. (11.10) is taken along C.

$$dQ = dq \int_c \gamma\, \nabla T \cdot d\mathbf{l} \tag{11.10}$$

where $d\mathbf{l}$ is an elementary vector on the curve C. The scalar $\nabla T \cdot d\mathbf{l}$ is simply equal to the differential dT, consequently

$$dQ = dq \int_{T_1}^{T_2} \gamma\, dT \tag{11.11}$$

This is the defining equation for the Thompson coefficient.

The Thompson relationships. The integral of Eq. (11.11) has the dimensions of voltage, and the terminal voltage of a thermoelectric loop such as that of Fig. 11.2 is due in part to the Thompson effect and in part to the Peltier

effect. The potential jumps at the junctions due to the Peltier effect are given by the expression

$$[\pi_1(T_2) - \pi_2(T_2)] - [\pi_1(T_1) - \pi_2(T_1)] \tag{11.12}$$

The voltage rise around the loop due to the Thompson effect in the two materials is

$$\int_{T_1}^{T_2} (\gamma_1 - \gamma_2)\, dT$$

Thus the total terminal voltage is

$$V = [\pi_1(T_2) - \pi_2(T_2)] - [\pi_1(T_1) - \pi_2(T_1)] + \int_{T_1}^{T_2} (\gamma_1 - \gamma_2)\, dT$$

Keeping T_1 constant and differentiating with respect to T_2 gives

$$\alpha = \frac{\partial V}{\partial T_2}\bigg)_{T_1 = \text{const}} = \frac{d(\pi_1 - \pi_2)}{dT} + \gamma_1 - \gamma_2 \tag{11.13}$$

In absolute terms then

$$\gamma_1 = \alpha_1 - \frac{d\pi_1}{dT} = \frac{\pi_1}{T} - \frac{d\pi_1}{dT} = -\frac{Td}{dT}\left(\frac{\pi_1}{T}\right) \tag{11.14}$$

and substitution from Eq. (11.8) gives

$$\frac{\gamma_1}{T} = -\frac{d\alpha_1}{dT} \tag{11.15}$$

Equations (11.7), (11.14), and (11.15) are the Thompson relationships which interrelate the various thermoelectric coefficients.

In the line integral of Eq. (11.10) it was tacitly assumed that the value of the line integral was independent of the path. This will normally be true only if the material is isotropic with respect to the Thompson coefficient. This assumption is carried through the entire discussion of the thermoelectric effect. We do not mean to imply that this is usually true of all materials. If we attempt to consider possible anisotropy of materials, we very quickly find ourselves in a maze of complex mathematics which is liable to obscure the points of major importance.

Absolute values of the thermoelectric coefficients. All of the thermoelectric coefficients of a superconductor are zero. Thus, if a junction is formed between a normal conductor and a superconductor, all of the thermoelectric effects will be due to the normal conductor alone and will be absolute properties of the single material. Because of this, it has been possible to measure the absolute Seebeck coefficient of lead up to 18°K.[3] Although it is

[3] J. W. Christian, J. P. Jan, W. B. Pearson, and I. M. Templeton, *Can. J. Phys.*, **36**, 627 (1958).

awkward, the Thompson coefficient of a material can be measured independently. This also has been done for lead down to 20°K.[4] Using these sets of data in combination with Eq. (11.15), it has been possible to evaluate the absolute Seebeck and Peltier coefficients of lead at any temperature.

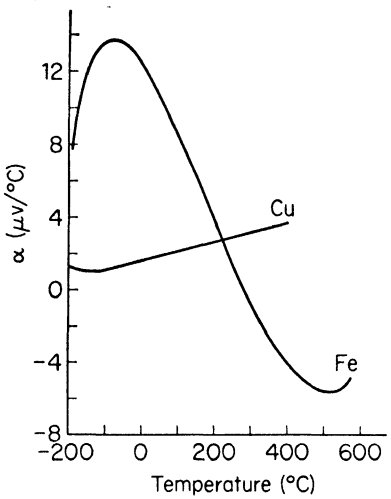


Figure 11.4 Absolute Seebeck coefficient of copper and iron as a function of temperature. (Domenicali and Otter, *J. App. Phys.* **26** (1955).)

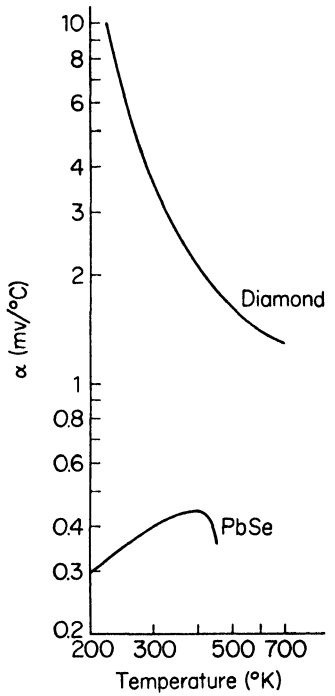


Figure 11.5 Thermoelectric power of semiconductors that have extremely different energy gaps. (For PbSe, Devyatkova and Smirnov, *Sov. Phys. Sol. State* **2**, 1786 (1961); for diamond, Goldsmid, Jenns, and Wright, *Proc. Phys. Soc.* **73**, 393 (1959).)

These in turn permit the determination of absolute thermoelectric coefficients for other materials.

In general, the Seebeck coefficients of metals are considerably lower than those of semiconductors. Figure 11.4 shows α for iron and copper over a range of temperatures. Figure 11.5 gives similar data for a semiconducting diamond and PbSe. Note that the units differ by a factor of 1000. Table 11.1

[4] G. Borelius, W. H. Keesom, C. H. Johansson, and J. O. Linde, *Proc. Acad. Sci. Amsterdam*, **35**, 10 (1932).

lists room temperature Seebeck coefficients for a number of metals and semiconductors.

Table 11.1. *Seebeck coefficients of certain materials at* 20°C*

Material	*mv/°C*
$ReSi_2$	174.0
$CrSi_2$	86.9
$TaSi_2$	14.0

* Neshpov and Samsonov, *Sov. Phys. Dok.*, **5**, 877 (1961).

11.3 The mechanism of thermoelectric phenomena

In any explanation of the underlying physical processes that are manifested as thermoelectric phenomena, the variation of the Fermi energy with temperature plays a leading role. Therefore it is desirable to digress somewhat before attacking the central problem, and to explore the factors that influence the variation of the Fermi energy with temperature.

Let us begin by considering a band whose density of states function is shaped as shown in Fig. 11.6 and which is half filled with electrons. At absolute zero, all of the states from the bottom of the band to the level marked aa' would be filled, and those above this level would be empty. Now as the temperature is increased, thermal excitation will cause electrons to leave states below the level aa' (which coincides with ζ) and enter states above the level marked aa'. Since the density of states function is, in this case, symmetric about aa' and since the Fermi-Dirac distribution function makes the probability of a state of energy $\zeta + U$ being occupied equal to the probability of a state of energy $\zeta - U$ being empty, the distribution of electrons in states above aa' will mirror the distribution of vacant states below aa'. Under these circumstances, the probability of a state at aa' being occupied will remain one-half regardless of the extent of thermal excitation and the Fermi energy will remain at aa' at all temperatures.

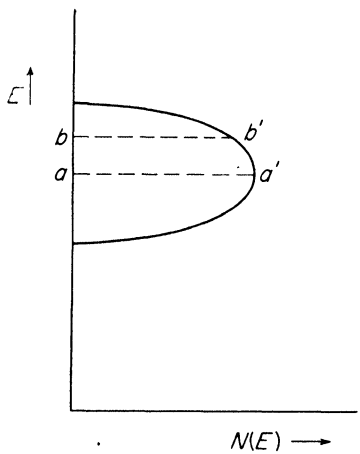


Figure 11.6 Density of states function in a hypothetical material. The line aa' indicates the level at which the band is half full, and the line bb' indicates the level when the band is more than half full.

If the density of states function of the band were not symmetric about the level of the Fermi energy at absolute zero, then the arguments presented above would no longer hold, and ζ would vary with temperature. For example, if the band considered above had the same density of states function, but

was 3/4 full, then at absolute zero all of the states below the line bb' would be occupied and those above bb' would be empty. If the temperature is increased, the distribution of the electrons above the Fermi level will have a lower value at any particular energy and consequently will extend over a broader range than would be true of the distribution of vacancies below the Fermi level. The probability of occupancy of a level at the Fermi energy would still be one-half, but this level would be somewhere below the level marked bb'.

In semiconductors this distribution is somewhat easier to visualize. In this case we are concerned with the distribution of electrons in more than one band. Consider an n-type semiconductor with the donor levels lying at an energy E_d, take the zero point of energy at the top of the valence band, and let the energy at the bottom of the conduction band be E_g. We shall assume further that thermal excitation will result in a transfer of electrons from the donor level to levels that are very close to the bottom of the conduction band. Under these circumstances, the Fermi level will coincide with E_d at absolute zero. As the temperature is increased, the Fermi level will move to a point halfway between E_d and E_g. As some of the electrons in the bottom of the conduction band are excited to higher levels in the conduction band, the Fermi level will continue to rise until excitation from the valence band becomes important. When this occurs, the Fermi level will start to drop and if the distribution of states near the top of the valence band is similar to that of the bottom of the conduction band, the Fermi level will approach $E_g/2$ as a limit as the temperature is increased.

The preceding comments provide a qualitative guide to the causes of the variation of the Fermi energy with temperature. In addition to this we require a means of comparing the absolute values of the Fermi energies of different solids at the same temperature. We may do this by selecting the zero point of the electron energy as its energy when it is outside of and well away from the surface of the solid in a field free region. If such an electron is then brought into the sphere of influence of the solid, it will, on the average, enter a state which is very near to the Fermi energy. In doing so, its energy will decrease by an amount ϕ, where ϕ is the Richardson work function. In an electrically neutral solid, then, the Fermi energy is the negative of the Richardson function. If the solid is not electrically neutral, the work function will vary with the absolute electrostatic potential of the material.

If two materials having different work functions are joined, their Fermi levels will differ initially, but as soon as contact is established between them, electrons will diffuse from the material having the higher Fermi energy to that having the lower Fermi energy. At equilibrium, a potential difference, $\phi_1 - \phi_2$, will exist between the two materials, and the Fermi energy of one will be depressed and that of the other elevated until they are equal. This is shown schematically in the sequence of Fig. 11.7.

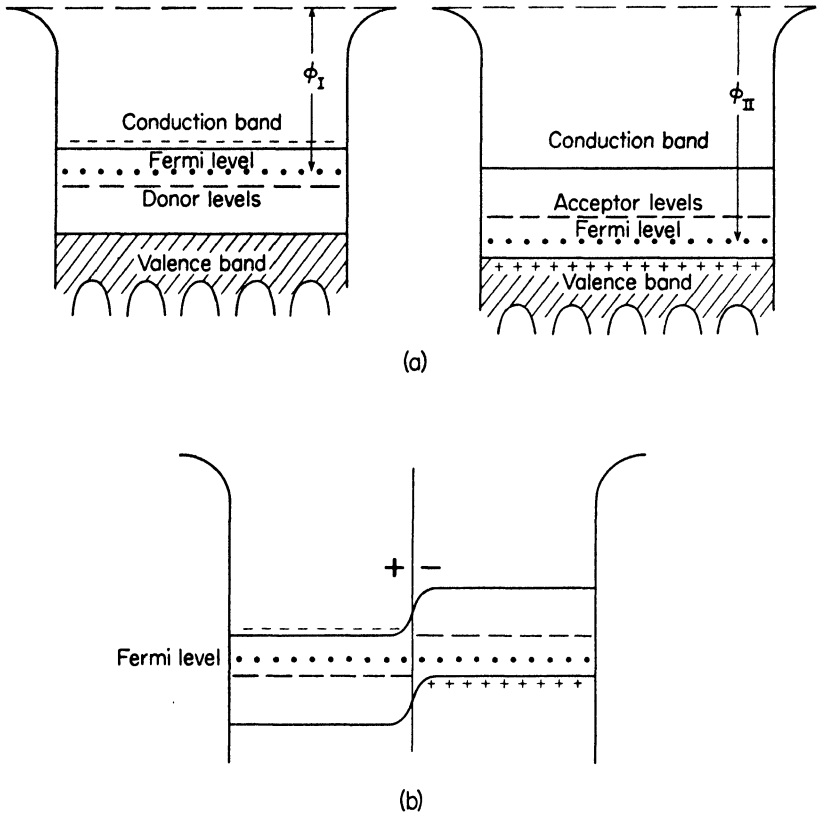


Figure 11.7 Fermi levels of two materials when separated and when joined and in equilibrium. Some of the holes have diffused into material I, have anniliated electrons there, and have given material I a net positive charge. Some of the electrons have diffused from material I into material II, have annihilated holes there, and have given material II a net negative charge.

If the two materials discussed above are joined at two different points to form a thermocouple, then the open-circuit terminal voltage will be

$$V = \phi_1(T_1) - \phi_2(T_1) + \phi_2(T_2) - \phi_1(T_2) \tag{11.16}$$

and differentiation gives

$$\alpha = \frac{d}{dT}(\phi_1 - \phi_2) \tag{11.17}$$

The absolute Seebeck coefficient of either material is therefore related to the work function by

$$\alpha_i = \frac{d\phi_i}{dT} = -\frac{d\zeta_i}{dT} \tag{11.18}$$

From Eqs. (11.8) and (11.15) we may write the relationships

$$\frac{\gamma_i}{T} = \frac{d^2\zeta}{dT^2} \quad \text{and} \quad \pi_i = -T\frac{\partial \zeta_i}{\partial T} \tag{11.19}$$

One of the obvious consequences of these relationships is that information about the variation of ζ with temperature can be obtained from thermoelectric measurements. This information can be combined with the results of other experiments to help determine the band structure of a solid.

The mechanism of the Peltier effect can be explained in terms of these ideas in the following way. If a charge q is passed through a boundary, it will either have to "climb" a potential hill equal to the contact potential, or will fall through a potential difference equal to the contact potential, depending upon the direction of passage. In the former case, it would have to abstract energy from the lattice, and heat would be absorbed at the junction; in the latter case, the charge would gain energy as it passed through the boundary. This would result in a liberation of thermal energy as the charge interacted with the lattice.

The Thompson effect is somewhat more difficult to visualize. In a region of the material sustaining a temperature gradient, the distribution of electrons and the concentration of electrons in any predetermined range of energies will vary in some manner along the temperature gradient. The resulting *concentration* gradient causes the electrons in this energy range to diffuse in a direction opposite to the concentration gradient.

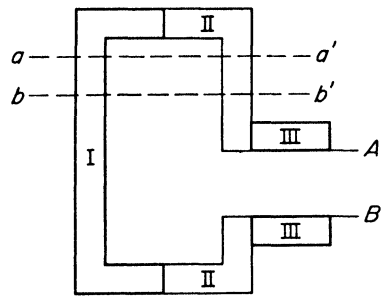


Figure 11.8 Thermoelectric circuit. Roman numerals indicate regions of different materials. The temperature gradient only exists in the materials between the lines aa' and bb'. Material 3 has no effect upon the terminal voltage of the circuit.

When we consider the entire distribution of electrons in energy along the temperature gradient, those electrons in one energy range might well be diffusing in one direction, while those in another range of energies could be diffusing in the opposite direction. Unless the rates of both processes coincide, there will be an accumulation of charge at one end of the temperature gradient. This in turn creates a potential gradient. In the steady state, the net transfer of charge will be zero because the combined effects of the concentration gradients and the potential gradient will balance one another.

In structures having cubic or higher symmetry, it is possible to deduce a relationship between the Thompson coefficient and the thermoelectric cross conductivity. In these cases, the thermoelectric cross conductivity becomes a scalar, which considerably simplifies the mathematics. Consider the circuit of Fig. 11.8. This consists of three different materials, two of which

are involved in the thermoelectric process, while the third simply provides electrical contact with the terminals. We distinguish three regions in the circuit which are labeled with Roman numbers. Material 1 occupies region I, region II is occupied by material 2, and region III by material 3. We impose the condition that everywhere above the line aa' the temperature is constant and has the value T_1, while everywhere below the line bb' the temperature is equal to T_2. We wish to express the open-circuit voltage in terms of the conductivities of the materials insofar as this may be possible.

Since the circuit is open, the electrical current density must be zero. Hence, from Eq. (11.1a) we may write

$$ {}_{e}\mathrm{C}\nabla V = -{}_{eT}\mathrm{C}\nabla T \tag{11.20}$$

In order to compute the terminal voltage, we may evaluate the line integral

$$ \int_C \nabla V \cdot d\mathbf{l} = -\int \frac{{}_{eT}\mathrm{C}}{{}_{e}\mathrm{C}} \nabla T \cdot d\mathbf{l} \tag{11.21}$$

along any curve connecting A and B and lying inside the materials, provided that V is a continuous function of the coordinates. In the case at hand, V is discontinuous at the boundaries between materials, and the potential jumps at these interfaces must be added to the line integral. If we include these potential jumps, we obtain

$$ V_{AB} = -\int_{T_2}^{T_1} \frac{{}_{eT}\mathrm{C}}{{}_{e}\mathrm{C}}\, dT + \int_{T_2}^{T_1} \frac{{}_{eT}\mathrm{C}'}{{}_{e}\mathrm{C}}\, dT + \pi(T_1) - \pi'(T_2) + \pi'(T_2) - \pi(T_1) \tag{11.22}$$

Here the primed terms refer to material 2 and the unprimed terms to material 1.

Keeping T_2 constant and differentiating with respect to T_1 results in

$$ \alpha = \frac{{}_{eT}\mathrm{C}'}{{}_{e}\mathrm{C}'} - \frac{{}_{eT}\mathrm{C}}{{}_{e}\mathrm{C}} + \frac{d}{dT}(\pi - \pi') \tag{11.23}$$

Notice that the Thompson coefficient has emerged as the ratio of the thermoelectric cross conductivity to the electrical conductivity. If we had adhered to tensor notation, γ, α, and π would all have emerged as rank four tensors. The tensor nature of the Peltier effect is best illustrated by the fact that a change in direction of the current density inside a single crystal anisotropic material such as bismuth telluride will either liberate or absorb heat just as though the current had crossed an interface between two different materials.

11.4 Applications of the thermoelectric effect

The use of thermocouples, thermopiles, and bolometers as temperature sensors in instrument systems has been exploited for many years. Such appli-

cations did not require any fundamental understanding of thermoelectricity, since only an empirical knowledge of the Seebeck coefficient would suffice. In the past seven or eight years, however, there has been considerable interest in the possible commercial exploitation of the Peltier effect in refrigeration and in the use of thermoelectricity in the direct conversion of heat to electricity. Certainly both of these may find highly specialized uses both in the general market and in technological applications. For example, Slovak peasants who have no access to electricity can buy a device capable of powering a radio which operates off the heat generated by a kerosene lamp.

A practical thermoelectric material for use in direct conversion of heat to electricity or in Peltier refrigeration must meet a host of stringent—and, to some degree, conflicting—requirements. Irreversible conduction of heat from the hot junction to the cold junction is the primary limitation of most thermoelectric materials. In applications such as these it is necessary to compromise between the requirements for high electrical conductivity and low thermal conductivity. Since the latter depends upon the transport of heat by both phonons and charge carriers, these requirements are to some degree mutually exclusive. However, phonon conduction can be minimized by choosing materials with complex lattices. The best materials for either of these applications at present are the semiconductors, Bi_2Te_3 and PbTe. Both of these may be made into either p- or n-type by adding impurities, or by synthesizing the crystals in such a way that there are controlled minor departures from stoichiometry. Consequently, both the legs of such a device may consist of lead telluride, for example, but one leg will be a p-type and the other an n-type semiconductor.

11.5 The Hall effect

When a charged particle moves through a magnetic field in free space, it is subject to the well-known Lorentz force

$$\mathbf{F} = e\mathbf{v} \times \mathbf{H} \tag{11.24}$$

This same force is operative upon the charge carriers conducting electricity through a solid, and is manifest as an electric field having a direction transverse to that of the electric current density vector. The magnitude and direction of this field is expressed in the equation

$$-\mathbf{E}_H = R\mathbf{H} \times \mathbf{J} \tag{11.25}$$

where R is called the *Hall coefficient.*

Assuming that only charge carriers of one sign are present

$$\mathbf{J} = ne\mathbf{v} \tag{11.26}$$

where $\mathbf{v}$ is the average drift velocity. Since by definition E_H is the force per

unit charge due to the Hall effect, Eqs. (11.24) through (11.26) become compatible only if

$$R = \frac{1}{ne} \tag{11.27}$$

Thus, in the case where only charge carriers of one sign are present, Hall coefficient measurements are capable of providing information about the concentration of charge carriers.

Since the presence of a magnetic field would otherwise perturb the current distribution in a conducting material, it is desirable to constrain the current density vector to a known direction through the entire sample when evaluating the Hall coefficient experimentally. This is accomplished by using a long thin specimen with cross arms for making connections to measure the Hall voltage. The normal sample geometry for making Hall measurements is shown in Fig. 11.9. A current is passed through the sample in the direction of its longitudinal axis by connecting the terminals aa' to a d-c power source. The magnetic field is applied in the x_3 direction of the figure and the Hall field is deduced from potential measurements between either the terminals bb' or cc'. The current density is assumed to be the total current divided by the cross-sectional area of the specimen. Simultaneous measurements of the resistivity can be made from potential measurements between terminals bc. Before the results of measurements made under these experimental conditions may be evaluated, we must be able to express the components of the resistivity tensor in terms of the coordinate system used to specify the experimental conditions. These depend upon the internal symmetry of the crystal rather than the external shape of the sample.

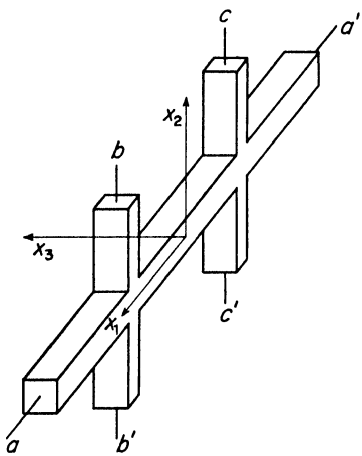


Figure 11.9 Sketch of sample and choice of coordinates for measuring the Hall effect.

For the most asymmetric class of crystals, the orientation of the principal axes of the resistivity tensor would have to be determined experimentally. However, by using Neumann's principle we can be assured that they will coincide with the crystallographic axes of crystals having orthorhombic or higher symmetry. Thus we can always choose to cut the crystal in such a manner that the coordinate axes of Fig. 11.9 will be principal axes of the resistivity tensor. Therefore the general expression for the potential gradient at a point in the material

$$-\nabla V = \boldsymbol{\rho} \cdot \mathbf{J} + R\mathbf{H} \times \mathbf{J} \tag{11.28}$$

would be subject to the conditions

$$J_2 = 0, \qquad J_3 = 0$$
$$H_1 = 0, \qquad H_2 = 0,$$
$$\rho_{ij} = 0, \qquad \text{for } i \neq j$$

Expanding Eq. (11.28) gives

$$-\nabla_1 V = \rho_{11} J_1$$
$$-\nabla_2 V = R H_3 J_1$$

where $-\nabla_1 V$ is the electric field intensity due to the d-c power source and $-\nabla_2 V$ is the Hall field intensity.

If the thermoelectric cross conductivity is zero,

$$\rho_{ij} C_{jk} = \delta_{ik} \qquad \text{and} \qquad \rho_{11} = \frac{1}{C_{11}}$$

Since $C_{11} = ne\mu_{11}$, where $\mathfrak{u}$ is the mobility tensor of the charge carriers, the ratio $\nabla_1 V / \nabla_2 V$ reduces to

$$\frac{\nabla_1 V}{\nabla_2 V} = \frac{1}{C_{11} R H_3} = \frac{ne}{\mu_{11} ne H_3} = \frac{1}{\mu_{11} H_3} \tag{11.29}$$

This is an important result since it indicates how the Hall field measurement combined with a conductivity measurement will provide information about the charge carrier mobility tensor.

If charge carriers of different signs are present, the Hall coefficient is more complex and it can be shown that the relationship

$$R = \left| \frac{1}{e} \right| \left\{ \frac{n^- \mathfrak{u}^- : \mathfrak{u}^- - n^+ \mathfrak{u}^+ : \mathfrak{u}^+}{(n^- \mathfrak{u}^- : \mathfrak{u}^- + n^+ \mathfrak{u}^+ : \mathfrak{u}^+ + 2n^- n^+ \mathfrak{u}^- : \mathfrak{u}^+)} \right\}$$

holds for these cases.

Figure 11.10a provides a practical illustration of the way Hall and resistivity measurements are combined to provide information about the mobility of charge carriers. The data presented there refer to n-type germanium. The raw Hall data are not shown. The charge carrier concentration was computed from the Hall voltage by Eq. (11.2).

Figure 11.10(b) shows Hall data for a typical n-type semiconductor. The sudden drop to zero indicates the onset of excitation of electrons from the valence band to the conduction band and is due to the fact that charge carriers of both signs are present in about equal concentrations. The subsequent variations in the Hall constant are caused by the variation of the mobilities of holes and electrons with temperature.

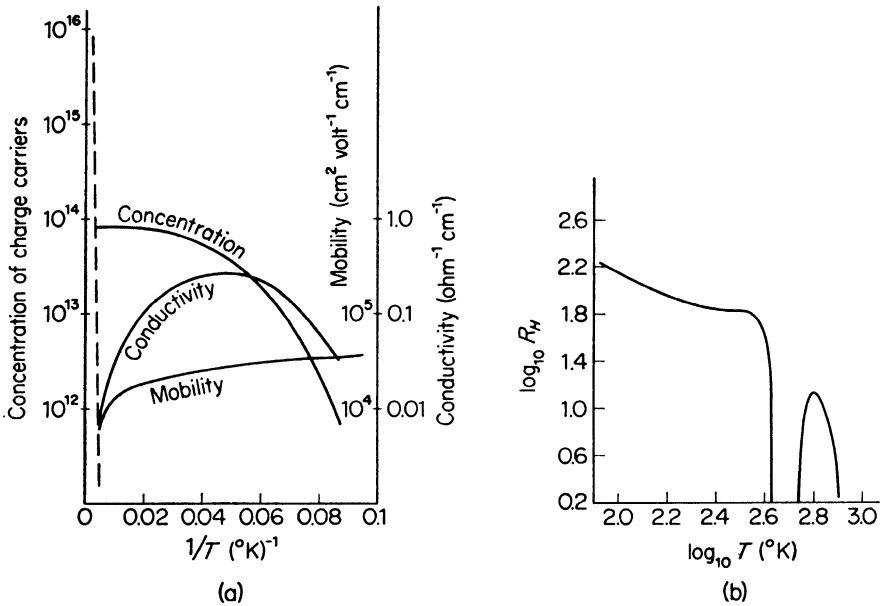


Figure 11.10 (a) Conductivity, concentration, and mobility of charge carriers in doped germanium. (From Kittel, *Intro. Solid State Phys.*, 2nd Ed., John Wiley & Sons, New York.) (b) Variation of R_H in Bi_2Te_3. (After Wooley and Gillett, *J. Phys. Chem. Solids* **17**, 34 (1960).)

11.6 Anisotropy of the Hall coefficient and magnetoresistance

Heretofore we have considered the Hall coefficient as a scalar. However, many single crystal materials are anisotropic with respect to their Hall coefficients; in order to account for this, R should be considered a tensor. The total interaction of such a material with an electric and a magnetic field is expressed in the tensor equation

$$-\nabla V = \varrho \cdot \mathbf{J} + \mathbf{R} \cdot (\mathbf{H} \times \mathbf{J}) \tag{11 30}$$

Expanding Eq. (11.30) and utilizing the Levi-Civita density to treat the cross product gives

$$-\frac{\partial V}{\partial x_k} = \rho_{kl} J_l + u_{lmn} J_n R_{kl} H_m \tag{11.31}$$

If ρ_{kl} and R_{kl} are functions of $\mathbf{H}$, as is often the case, we may write the following approximate relationships for weak magnetic fields

$$\rho_{kl}(\mathbf{H}) = \rho_{kl}(\mathbf{H} = 0) + \frac{\partial^2 \rho_{kl}}{\partial H_m \, \partial H_n} H_m H_n + \ldots$$

$$R_{kl}(\mathbf{H}) = R_{kl} + \frac{\partial^2 R_{kl}}{\partial H_n \, \partial H_o} H_n H_o + \ldots$$

The quantities

$$\frac{\partial^2 \rho_{kl}}{\partial H_m \, \partial H_n} = \rho_{kl,mn} \qquad \text{and} \qquad \frac{\partial^2 R_{kl}}{\partial H_n \, \partial H_o} = R_{kl,no}$$

are rank four tensor *galvanomagnetic* properties of the material. Incorporating these into Eq. (11.31) gives

$$-\frac{\partial V}{\partial x_k} = \rho_{kl} J_l + \rho_{kl,mn} J_l H_m H_n + u_{lmn} R_{kl} H_m J_n + u_{lmn} R_{kl,no} H_m H_n H_o J_n \tag{11.32}$$

In the case of the experimental set-up shown in Fig. 11.9, the only component of $\mathbf{H}$ is H_3 and the only component of $\mathbf{J}$ is J_1. The field in the x_1 direction then reduces to

$$-\frac{\partial V}{\partial x_1} = \rho_{11} J_1 + \rho_{11,33} J_1 H_3^2$$

The last terms disappear because $n = l$ making the Levi-Civita density zero.

If we allow the direction of $\mathbf{H}$ to vary, but confine it to the $x_1 x_2$ plane, then

$$-\frac{\partial V}{\partial x_1} = \rho_{11} J_1 + \rho_{11,12} H_1 H_2 J_1$$

The second term describes the *magnetoresistance effect*, and is usually measured as a change in resistivity divided by the zero field resistivity

$$\frac{\Delta \rho}{\rho} = \frac{\rho_{11,12} H_1 H_2}{\rho_{11}} \tag{11.33}$$

The mechanism of magnetoresistance is described in the following paragraphs.

If an electric field is applied to a solid in some direction other than that of a principal axis of the resistivity tensor, the current will have a direction that does not coincide with that of the negative of the electric field. Now if a magnetic field is applied in the same direction as the electric field, the transverse components of the current will be deflected and will begin to circulate around an axis oriented with the applied electrical and magnetic fields. The circulating components of the current will in turn produce a magnetic field which augments the applied magnetic field.

The net result of this is that the circulation of the transverse components of the current density depends upon the square of the applied magnetic field and the square of the charge carrier mobility. In the absence of a magnetic field, the current streamlines will be straight lines making some angle with the applied field. When the magnetic field is introduced, the current streamlines become helices and the current must travel a path which may be many times as great as the direct path through the material from one electrode to the other. The result is an increase in resistance as measured between the terminals of the sample. Figure 11.11 shows magnetoresistance data for InSb.

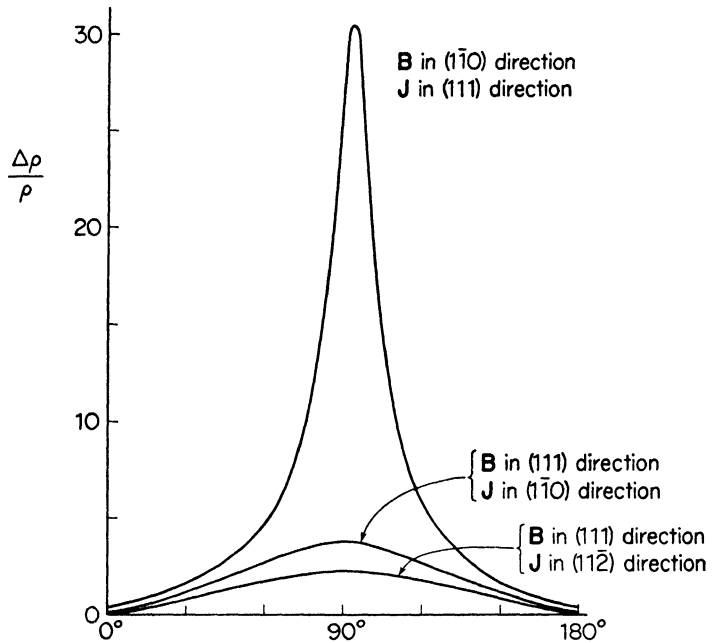


Figure 11.11 Magnetoresistance in InSb at 295°K. (Rupprecht, Weber, and Weiss, *Z. Naturforsh.* **15a**, 783–784 (1960).)

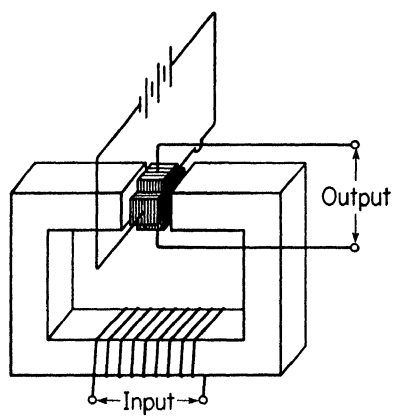


Figure 11.12 The circuit shown could be used as a Hall amplifier or as a device capable of performing a multiplication operation.

Both the Hall effect and magnetoresistance are important in determining the band structure of the solid since they are capable of providing information about the concentration and mobility of the charge carriers and the shape of the energy surfaces in the solid. These phenomena also have practical applications in devices. The most obvious application of magnetoresistance is that it provides a means of varying resistance without using a sliding contact. Two applications of the Hall effect are illustrated in Fig. 11.12 which shows a circuit diagram of a device which can act either as a power amplifier or as an element having an output voltage proportional to the product of two input currents.

11.7 Thermomagnetic and thermogalvanic effects

Since moving charges transport thermal energy as well as electrical charge and interact with a magnetic field without regard to which quantity is being transported, we might well inquire about the entire spectrum of phenomena which might result from the totality of possible interactions between a temperature gradient, an electric field, and a magnetic field. In the most general case, these three quantities, the electric current density, and the entropy current density would be related by the set of equations

$$\begin{aligned} -\nabla V &= {}_{e}\varrho \cdot {}_{e}\mathbf{J} + {}_{es}\varrho \cdot {}_{s}\mathbf{J} + \mathbf{R} \cdot (\mathbf{H} \times {}_{e}\mathbf{J}) + \mathbf{S} \cdot (\mathbf{H} \times {}_{s}\mathbf{J}) \\ -\nabla T &= {}_{es}\varrho \cdot {}_{s}\mathbf{J} + {}_{s}\varrho \cdot {}_{s}\mathbf{J} + \mathbf{S} \cdot (\mathbf{H} \times {}_{e}\mathbf{J}) + \mathbf{L} \cdot (\mathbf{H} \times {}_{s}\mathbf{J}) \end{aligned} \tag{11.34}$$

We identify the terms containing the cross products as:

1. $\mathbf{R} \cdot (\mathbf{H} \times {}_{e}\mathbf{J})$—the Hall field
2. $\mathbf{S} \cdot (\mathbf{H} \times {}_{s}\mathbf{J})$—the Nernst field
3. $\mathbf{S} \cdot (\mathbf{H} \times {}_{e}\mathbf{J})$—the Ettinghausen temperature gradient
4. $\mathbf{L} \cdot (\mathbf{H} \times {}_{s}\mathbf{J})$—the Leduc-Righi temperature gradient

The significance of each is apparent from its appearance in Eqs. (11.34), i.e. $\mathbf{S} \cdot (\mathbf{H} \times {}_{s}\mathbf{J})$ is a transverse electric field developed from the interaction of an entropy current and a magnetic field.

The quantities **R**, **S**, and **L** are all rank two tensor properties of the material under consideration. Note that the same tensor characterizes the Nernst and the Ettinghausen effects. This reciprocity can be proven by the methods of irreversible thermodynamics.

The De Haas–van Alphen effect and related phenomena. In Sec. 11.6 we noted that the components of the Hall coefficient R_{ij} are functions of the magnetic field intensity. Normally Hall data in the literature are interpreted as the zero field intensity values of R_{ij}. At intermediate fields the variation in the components R_{ij} can be expressed in terms of a Taylor expansion, and in

this form we can define a tensor $\partial^2 R_{ij}/\partial H_k \partial H_l$ which is used in characterizing the magnetoresistance effect. At higher magnetic fields, a rather remarkable phenomenon occurs: the components of **R** begin to vary cyclically with increasing field as shown in Fig. 11.13. This is known as the *De Haas-van Alphen effect.* The components of the Nernst effect tensor and the Leduc-Righi tensor will undergo similar cyclic variations with increasing magnetic

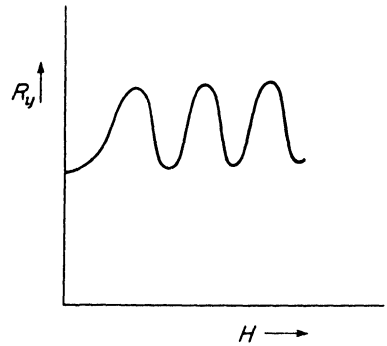


Figure 11.13 Variation of the Hall coefficient with magnetic field intensity illustrating the de Haas–van Alphen effect.

field intensity. These phenomena are important in investigations of the shape of the Fermi surface, and are of considerable interest from a theoretical point of view; however, a consideration of their mechanisms is beyond the scope of this book.

PROBLEMS

11.1 Two metals labeled 1 and 2 are incorporated into a thermoelectric circuit. One junction is held at 0°C. The terminal voltage as a function of temperature of the other junction is tabulated below. The absolute Seebeck coefficient of metal 2 is also tabulated. Find the absolute values of α, γ and π for metal 1 at 30°C.

T °C	V	α_2
0	0	
10	0.030	0.00050
20	0.058	0.00048
30	0.086	0.00046
40	0.112	0.00044
50	0.137	0.00042
60	0.162	0.00040

11.2 If the electrical conductivity of metal 1 in the previous problem is 10^6 ohm^{-1}cm^{-1} and the metal is isotropic, compute the thermoelectric cross conductivity and the electrical resistivity of the metal.

11.3 Predict qualitatively the sort of variation with temperature that would be expected of the absolute Seebeck coefficient for an n-type semiconductor in the temperature ranges in which the following processes are occurring:

(a) Excitation from the valence band is negligible, 10 per cent of the donors are ionized.

(b) Excitation from the valence band is negligible, 90 per cent of the donors are ionized.

(c) Practically all of the donors are ionized and excitation from the valence band is just becoming appreciable.

11.4 The following values represent the components of the electrical, thermal, and thermoelectric conductivities of a hexagonal material. Compute the electrical resistivity of the material. The units are not given; assume that the members represent a self-consistent set of units.

$$
\begin{aligned}
&{}_{e}\mathbf{C}_{12} = {}_{e}\mathbf{C}_{13} = {}_{e}\mathbf{C}_{23} = 0 \qquad {}_{e}\mathbf{C}_{11} = {}_{e}\mathbf{C}_{22} = 2 \qquad {}_{e}\mathbf{C}_{33} = 5 \\
&{}_{eT}\mathbf{C}_{11} = {}_{es}\mathbf{C}_{22} = 0.01 \qquad {}_{T}\mathbf{C}_{11} = {}_{T}\mathbf{C}_{22} = 0.05 \qquad {}_{T}\mathbf{C}_{33} = 0.12 \\
&\qquad\qquad\qquad\qquad {}_{eT}\mathbf{C}_{33} = 0.02
\end{aligned}
$$

11.5 A planar single crystal of a hypothetical material has the following properties:

$$
\begin{aligned}
R_{11} &= -17 \times 10^{-11} \text{ volt-cm/amp. Oersted} \\
R_{22} = R_{33} &= -30 \times 10^{-11} \text{ volt-cm/amp. Oersted} \\
{}_{e}\rho_{11} &= 10^{-4} \text{ ohm-cm} \\
{}_{e}\rho_{22} &= 2 \times 10^{-4} \text{ ohm-cm} = {}_{e}\rho_{33}
\end{aligned}
$$

The electrical current density at a point in the sample has the components: $J_1 = 0.5$ amp., $J_2 = 1$ amp., and a magnetic field whose components are $H_1 = H_2 = 0$, $H_3 = 10^4$ Oersted. Compute the potential gradient in the material.

11.6 Magnetoresistance measurements are made upon a single crystal of a cubic material whose electrical resistivity is 10^{-3} ohm-cm. Choosing a coordinate system aligned with the crystallographic axes, a current density of $J_1 = 10$ amp./cm^2 is supplied and is held constant. A magnetic field of 5000 Oersteds is present and is parallel to the x_1x_2 plane. The magnetic field can be rotated, and the following data relating the potential gradient in the crystal to the angular position of the magnetic field with respect to the x_1 axis is obtained:

$-\partial V/\partial x_1$	Θ	$-\partial V/\partial x_2$
1.010 volts/cm.	0°	0
0.882	30°	0.108
0.635	45°	0.125
0.635	60°	0.108
0.510	90°	0

Evaluate the components $\rho_{11,11}$, $\rho_{11,22}$, and $\rho_{21,21}$ of the magneto-resistance tensor.

12

DIELECTRIC AND MAGNETIC PROPERTIES

Physicists and philosophers have long been bemused by the fact that objects exert forces upon one another through apparently empty space, and it is vaguely disturbing to find that there are three distinct types of field through which this action at a distance occurs. All physical objects exert a gravitational attraction upon one another which varies as the inverse of the square of the distance of separation and is proportional to a quality we call *mass*. Many, but not all, fundamental particles also have a quality called *charge*. These exert an additional force upon one another which also varies according to the inverse-square law, but here the analogy ceases. There are two different kinds of charge, and the direction of the force that two particles exert upon one another depends upon their charges. The situation becomes more confusing when we consider that moving charges give rise to a third field, the magnetic field, whose sense depends upon both the charge and the direction of motion.

The fact that electric and magnetic fields both ultimately have their source in charged particles confers a certain kinship upon them that excludes the gravitational field and leaves it in a different class. This kinship carries over into the response that a material will exhibit when exposed to either an electric field or a magnetic field, and we shall find that there is a close similarity of form in the equations used to describe dielectric and magnetic behavior of solids. Many of the concepts that we shall wish to develop for a description of either of these also have an analogue in the description of the other. Consequently, it is desirable to treat both topics in the same chapter.

12.1 Polarization in a dielectric

When a dielectric is exposed to an electric field, the field exerts forces upon the charged fundamental particles comprising the dielectric. These forces act in opposite directions upon electrons and nuclei, positive and negative ions, and upon the ends of molecules having permanent dipole moments. The result is that the electric field causes a displacement of the centroid of the negative charge distribution in the system relative to the centroid of the distribution of positive charges. This is called *polarization.* That component of polarization due to the displacement of electrons and nuclei is called *electronic polarization,* and the other mechanisms indicated are called respectively *ionic polarization* and *dipole orientation.* Any or all of these may be operative at the same time, depending upon the nature of the dielectric and the frequency of the applied field. There are also a number of variations upon these basic mechanisms which we shall not concern ourselves with here.

In addition to the polarization induced by an electric field outside the dielectric, there are some crystalline materials that polarize spontaneously under the action of fields arising from sources inside the system. In any case, the total electric moment of the system is a vector parameter which can be expressed formally in the equation

$$\mathbf{m} = \Sigma\, q_i \mathbf{s}_i \tag{12.1}$$

where q_i is the charge on the ith particle and $\mathbf{s}_i$ is its displacement relative to a coordinate system fixed in the system.

The polarization per unit volume is defined by the equation

$$\mathbf{P} = \frac{\mathbf{m}}{V} \tag{12.2}$$

P is an extensive parameter and as such will be incorporated into the definition of the dielectric rigidities and compliances.

The external field intensity. The electric field intensity may be regarded as an intensive system parameter and as such it, too, will be involved in the definition of the dielectric properties of solids. The local field at a point inside a dielectric is a composite of a field arising from the polarization of the surrounding medium and a field arising from sources outside the dielectric. *It is only this latter portion of the field* which is considered to be an intensive system parameter. We refer to this as the "external" field although the term is something of a misnomer. Even though it is actually present inside the dielectric as a composite part of the local field, it is characterized in terms of a distribution of charges outside the dielectric.

The characterization of this external field is accomplished by first considering two parallel plane sections having the same lateral dimensions and

area A separated by a distance l. A charge q is removed from one plane and placed upon the other. This gives each plane an equal charge but the charges are opposite in sign. The charge is uniformly distributed over each plane. If there is no material between the planes and if l is small in comparison with the lateral dimension of the plane sections, the electric field intensity between these will be

$$E = \frac{q}{K^\circ A l} \tag{12.3}$$

where K° is the permittivity of free space.

When a dielectric is interposed between the planes, completely filling the region between them, the field will induce a polarization in the dielectric in such a direction that the induced polarization will neutralize a portion of the charge on the planes. We shall designate that portion of the charge which is neutralized by the polarization as q_b and the remainder as q_f. The total charge then is

$$q = q_b + q_f \tag{12.4}$$

The interposition of the dielectric between the planes also reduces the field intensity. Originally the total charge q produced the field but after a portion of this is neutralized, the field reduces to

$$E = \frac{q_f}{K^\circ A} \tag{12.5}$$

In order to relate P and q_b, we need only note that the term "neutralize" implies that the moment of q_b is equal and opposite to the moment of the dielectric. Thus

$$A\mathbf{P} \cdot \mathbf{l} = q_b \mathbf{l} \tag{12.6}$$

The dielectric displacement. If we compute the total charge displacement including both the polarization displacement **m** and the displacement of charge causing the external field, the result is a vector called the *dielectric displacement*. The dielectric displacement per unit volume is

$$\mathbf{D} = \mathbf{P} + \frac{q_f \mathbf{l}}{Al} \tag{12.7}$$

Combining this with Eq. (12.6) gives

$$\mathbf{D} = \mathbf{P} + \frac{K^\circ E \mathbf{l}}{l} = \mathbf{P} + K^\circ \mathbf{E} \tag{12.8}$$

12.2 Dielectric rigidities and compliances

In general, the polarization in a material will be a function of stress, temperature, and electric field intensity. The compliances used to characterize the

response of the material are the partial derivatives in the equation

$$d\mathbf{P} = \frac{\partial \mathbf{P}}{\partial T}\Big)_{\sigma,E} dT + \frac{\partial \mathbf{P}}{\partial \boldsymbol{\delta}}\Big)_{T,E} : d\boldsymbol{\delta} + \frac{\partial \mathbf{P}}{\partial \mathbf{E}}\Big)_{T,\sigma} \cdot d\mathbf{E}$$

The first two of these, the pyroelectric coefficient and the piezoelectric stress coefficient, are treated in subsequent sections. The last is called the *dielectric susceptibility* and is designated as χ_e

$$\chi_e = \frac{\partial \mathbf{P}}{\partial \mathbf{E}}\Big)_{T,\sigma} \tag{12.9}$$

The dielectric susceptibility is a rank two tensor property and as such is subject to the usual restrictions imposed by Neumann's principle. We could also define the corresponding rigidity as $\partial \mathbf{E}/\partial \mathbf{P}_{s,\varepsilon=\text{const}}$, but this is used so rarely that it has not been given a name.

We note that when Eq. (12.8) is differentiated, we have the result

$$\frac{\partial \mathbf{D}}{\partial \mathbf{E}}\Big)_{T,\sigma=\text{const}} = \chi_e + \mathbf{K}^\circ \tag{12.10}$$

In order to make Eq. (12.10) dimensionally homogeneous, the permittivity of free space must be regarded as a rank two tensor. Since free space is isotropic, $\mathbf{K}^\circ$ will be diagonal in any coordinate system. The sum

$$\mathbf{K} = \chi_e + \mathbf{K}^\circ \tag{12.11}$$

is called the *dielectric permittivity* of the material, or more frequently the *dielectric constant.*

K is significant in that it provides a means of comparing the capacitance of geometrically identical capacitors having different dielectrics between the plates. We can illustrate this by comparing the charges that would reside on the plates of two geometrically identical capacitors having the same potential difference between the plates but different dielectric materials between the plates. For simplicity, we choose an isotropic dielectric material between the plates of one capacitor and leave the space empty between the plates of the other. Then, the susceptibility and the permittivity will be scalars rather than tensors, and the direction of E and P will be perpendicular to the plates.

If E and P are linearly related and if $\mathbf{P} = 0$ when $\mathbf{E} = 0$, then

$$D = \frac{ql}{Al} = EK$$

for either capacitor. This can be rewritten as

$$K = \frac{ql}{AV}$$

where V is the potential difference between the plates. Therefore, the relative

charges on the plates of the two capacitors under consideration will be

$$\frac{K}{K^\circ} = \frac{q}{q_f} \tag{12.12}$$

Since they differ only by the additive term K°, the dielectric constant and the dielectric susceptibility are governed by the same fundamental processes operating in the material. In general, the dielectric susceptibility is of most interest in cases where the objective is to obtain information about these fundamental processes, while the dielectric constant is a more useful quantity for designing electronic components or for analyzing the action of these components in a device.

12.3 The relaxation time

It will be recalled that polarization can occur through several mechanisms the most important of which fall into the categories called electronic polarization, ionic polarization, and dipole orientation. In each of these cases, particles having a nonzero mass are displaced to new equilibrium positions in

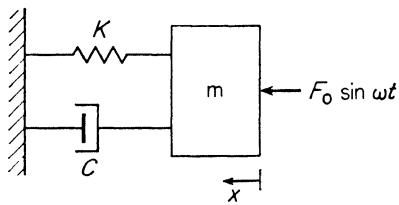


Figure 12.1 Mass-spring–dashpot system. The response of this system to a time-dependent force is analogous to the way that the polarization in a dielectric responds to a time-dependent electric field.

response to forces exerted upon them by the electric field. Just as is the case for all material objects, these are subject to Newton's laws of motion and when the electric field changes with time, their inertial reaction will introduce a time lag between changes in the electric field and the resulting displacement of charge. In addition to this, the motion of a particle relative to its neighbors will also involve work being done against local fields, and a portion of this may be transferred to the lattice as thermal energy instead of being stored as potential energy.

Although it is not desirable to carry the analogy too far, there are many qualitative similarities between the behavior of the polarization as a function of frequency of the electric field, and the amplitude of the response of a mass-spring–dashpot system subjected to a sinusoidal force. Consider the system shown in Fig. 12.1. The differential equation of motion is

$$m\ddot{x} + c\dot{x} + kx = F_0 \sin \omega t$$

This may be rewritten in the form

$$\tau^2\ddot{x} + 2\xi\tau\dot{x} + x = X_0 \sin \omega t \tag{12.13}$$

where τ is called the *relaxation time*, ξ the damping ratio, and X_0 the amplitude of the zero frequency displacement. The solution of Eq. (12.12) leads to a steady-state response amplitude

$$X = X_0 \left\{ (\tau^2\omega^2 - 1)^2 + \frac{4\xi^2\omega^2}{\tau^2} \right\}^{-1/2} \tag{12.14}$$

The tangent of the phase angle between the displacement and X_0 is equal to the ratio of the energy dissipated to the energy stored per cycle:

$$\tan \delta = \frac{2\xi\omega}{\tau(\tau^2\omega^2 - 1)} \tag{12.15}$$

The general features of the response are shown in Fig. 12.2a for $\xi < 1$. Notice that for $\omega > 1/\tau$ the response amplitude falls off rapidly with increasing frequency, and that at resonance, a large proportion of the energy supplied to the system is dissipated rather than stored. (See Figure 12.2b.)

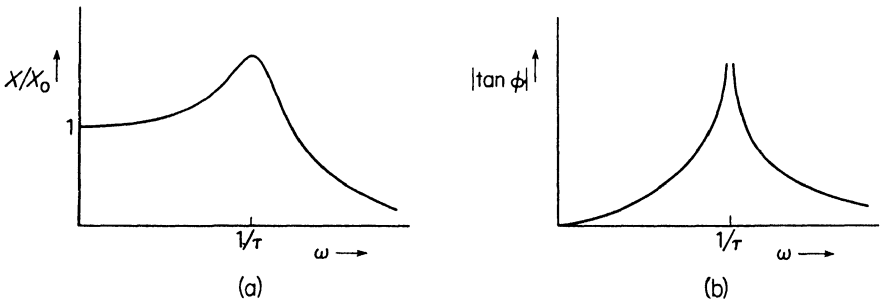


Figure 12.2 (a) Amplitude of response as a function of frequency for an underdamped mass-spring–dashpot system. (b) Absolute value of the tangent of the phase angle between the driving force and the response of a mass-spring–dashpot system.

The response of an ion or an electron in a solid to the force exerted by the electric field is quite similar to that of the mass in the mechanical system to the driving force. The chief difference is that the dissipative forces experienced by a charged particle in a solid are not represented very well by the dashpot. Nevertheless, the variation in the dielectric constant with frequency and the variation of the dielectric losses are very similar to the curves of Fig. 12.2. Where several polarization processes are operative, they all respond to slow variations in the field. If the field is varied sinusoidally, then as the frequency is increased, a point is reached where the polarization process having the longest relaxation time can no longer follow the reversals in the

field. When this happens, it ceases to contribute to polarization, the loss angle goes up, and the dielectric constant falls off. As the frequency increases further, such points will be reached for the processes having successively shorter relaxation times. Figure 12.3 is typical of the variation of dielectric constant in which the three distinct polarization processes are operative.

Evidently the dielectric constant is singularly lacking in constancy; in a single crystal dielectric it is a tensor and varies with crystallographic direction. Even in isotropic materials it varies with the frequency of the applied field; moreover it is strongly temperature dependent.

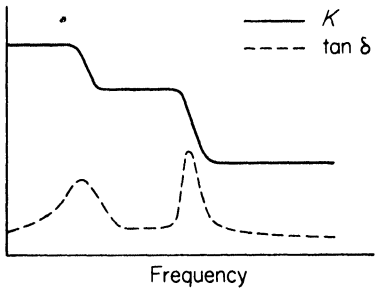


Figure 12.3 Variation of K and tan δ with frequency for a typical dielectric.

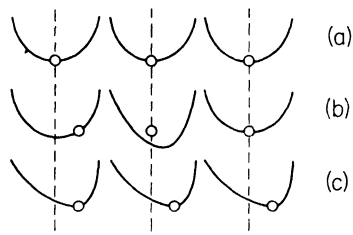


Figure 12.4 Schematic representation of polarization mechanism.

12.4 Spontaneous polarization in a crystal

We mentioned in a previous section that some crystalline materials become polarized spontaneously under the action of internal fields. Normally this only occurs concurrently with a phase transformation as the crystal is cooled below some critical temperature above which it is not polar. In this section we wish to explore the way a material might undergo a structural change leading to a spontaneous polarization.

Each ion in a crystal lattice exists in a force field consisting of electrostatic forces and short-range repulsion forces. The lattice, of course, will exist in a structure in which the total potential energy derived from these force fields is at a minimum. As the temperature changes, these force fields change, largely because of dimensional changes in the lattice; hence the structure may become unstable with respect to a phase transformation.

Let us select one ion in the basic cell of a structure in which the centroid of the negative charges coincides with that of the positive. As the temperature is reduced, it will ultimately reach a value at which the potential energy surface associated with the ion will be flattened in the neighborhood of its equilibrium position. This is shown in Fig. 12.4a. In the figure only one ion in a series of adjacent basic cells is depicted. The other ions in the lattice are presumed to be fixed in position. Now if ion 1 is displaced to the right

by a thermal transient, it will warp the potential surface around ion 2 as shown in Fig. 12.4b. When ion 2 moves to the potential minimum, it will disturb the potentials around ions 1 and 3, so that they too will move to the right. The net result is that all of the ions on that line will be displaced to the right relative to the other ions in the lattice. Since the ions of this kind are all of the same sign and are displaced relative to all of the other ions in the lattice including those of opposite sign, a polarization accompanies this structural change. Thus, the initial displacement of ion 1 triggers a polarization catastrophe that spreads very rapidly in the direction in which it was initiated.

This process will also disturb the potential distribution around ions residing in basic cells on lines parallel to the line upon which it was initiated. The result is that, once begun, the polarization catastrophe will spread through the crystal in a transverse direction also, but the rate at which it

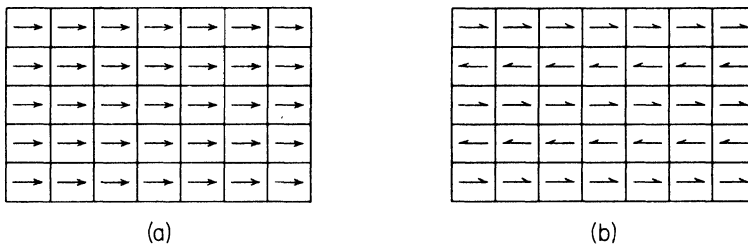


Figure 12.5 Schematic representation of parallel and anti-parallel polarization—each square represents a basic cell.

spreads in this direction is lower than the rate at which it proceeds in the polarization direction.

Spontaneous polarization will not occur in all ionic materials since it requires a local distribution of interionic forces such that the potential field around one of the ions in each basic cell becomes rather flat in the neighborhood of the original equilibrium position of the ion, and that a small displacement of this ion will exert an appreciable electrostatic force upon the corresponding ion in the next basic cell. This combination of circumstances is not particularly rare, but it is the exception rather than the rule.

Spontaneous polarization may have the result that all basic cells of a region in a crystal are polarized in the same direction, or the result may be that adjacent parallel rows of basic cells are polarized in opposite directions. The former situation is called *parallel polarization*, the latter *antiparallel polarization*. (See Fig. 12.5.)

The Curie temperature. The transition temperature at which spontaneous polarization occurs is called the *Curie temperature*. Above the highest Curie

temperature, the structure will have either a center of symmetry or some other combination of symmetry elements such that the centroid of the positive and negative ions of each basic cell coincide. When the material is cooled below the highest Curie temperature, it undergoes a phase transition to an asymmetric polarized structure. In most cases, the polarized structure is derived from the nonpolarized structure by relatively small dimensional changes in the basic cell. The direction of spontaneous polarization will have some simple relationship to the crystallographic axes of the nonpolarized structure. For example, if a cube, the spontaneous polarization will often be along an edge, a face diagonal, or a body diagonal. Any particular material may have a number of Curie temperatures corresponding to transitions from a nonpolarized or paraelectric state to a polarized state, from a polarization on an edge direction to a face or body direction, or between a parallel and an antiparallel state.

Polarization domains. When a single crystal is cooled through its Curie temperature, the polarized phase may be nucleated at several different points in the crystal. Since there may be a number of equivalent crystallographic directions in which the spontaneous polarization can occur, these

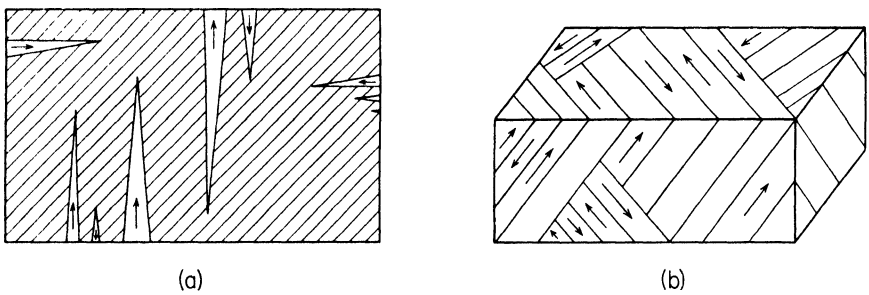


Figure 12.6 (a) Schematic of domain growth into a paraelectric phase. (b) Schematic of the domain structure in a single crystal.

nuclei will normally differ in polarization direction. Thus, as they grow through the crystal, they form several regions or domains differing in their direction of polarization.

Polarization involves some distortion of the basic cell, and the domain walls will consequently be in a state of strain. However, the dimensional changes accompanying polarization are relatively small, and even though the domain wall marks an interruption in the regularity of the lattice, these walls are not regarded as grain boundaries between different crystals. Instead, the domain is regarded a subgrain within a single crystal.

When the polarized phase is first nucleated, it grows much more rapidly in the direction of polarization than in transverse directions. For this reason,

the growing domains are usually wedge-shaped. Thus during a transition from a nonpolarized state to a polarized one, the domain structure might appear as shown in Fig. 12.6a. The crosshatched area represents that portion of the crystal which has not yet been transformed to the polarized state. Figure 12.6b shows a schematic of the domain structure after the transformation is complete.

12.5 Pyroelectric materials

Once a material has undergone a polarization catastrophe, one might expect a large potential difference to exist between opposing faces perpendicular to the polarization direction. This is detectable only when the faces are freshly cut, since the polarization is rapidly neutralized by electrical leakage and the accumulation of stray charges upon the surfaces. Historically a manifestation of the internal polarization was first noticed as a transient electric field following a temperature change, hence the name *pyroelectricity*. Because the internal polarization is temperature dependent, a temperature change produces a temporary imbalance between the internal electric moment and that due to the accumulated surface charges.

12.6 Ferroelectric materials

The direction of polarization in some pyroelectric materials can be altered by an external electric field. These materials are called *ferroelectric*. These are divided into four classes upon the basis of structure: Rochelle salt and its isomorphs; KH_2PO_4 and its isomorphs; guanidine and its isomorphs; and those crystals having the perovskite or the ilmenite structures in their paraelectric phases. Certain of these materials are listed with their Curie temperature in Table 12.1.

Table 12.1 *Certain ferroelectric crystals*

Crystal	*Structure*	*Curie temperature* (°*K*)
$NaK(C_4H_4O_6) \cdot 4H_2O$		297 (upper)
		255 (lower)
$KNbO_3$	perovskite	708
$PbTiO_3$	perovskite	763
$LiTaO_3$	ilmenite	
KH_2PO_4	ilmenite	213
RbH_2AsO_4		111
CsD_2AsO_4		212

We cannot treat all of these in detail, and shall direct most of our remarks toward barium titanate which has the perovskite structure. In passing, however, we remark that the polarization catastrophe in the Rochelle salt class involves displacement of the hydrogen atoms in the water of hydration, and

that in the KH_2PO_4 class it involves the shift of hydrogen atoms in hydrogen bonds. The mechanism of polarization in the guanidine class is still a matter of controversy.

The structure of barium titanate in the paraelectric phase is shown in Fig. 12.7. The polarization mechanism is fairly well established and consists of a displacement of the titanium ions relative to the surrounding oxygen octahedra. The highest Curie temperature is 120°C., and the transition to the ferroelectric state involves a polarization in the direction of an edge of the

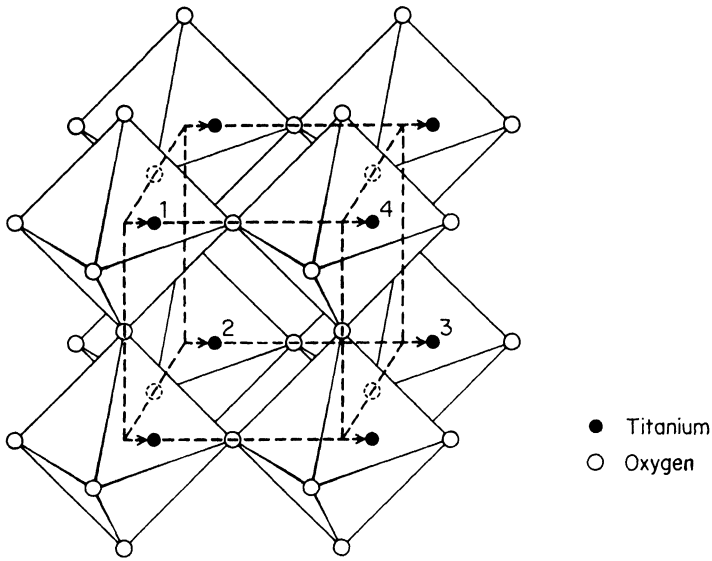


Figure 12.7 $BaTiO_3$ structure in polarized form. The barium ion is not shown. It is located at the center of the cube. Each corner is surrounded by octahedral oxygen ions. The polarization arises from a displacement of the titanium ions relative to the oxygen octahedra. In the paraelectric phase, these are located at the corners of the cube. In the anti-ferroelectric perovskites, the numbered titanium ions are displaced in opposite directions from the un-numbered titanium ions.

original cubic structure. The structure elongates in the direction of polarization and suffers a transverse contraction. Two Curie temperatures occur at 0°C and −80°C. These mark a shift of the polarization to that of a face diagonal and a body diagonal respectively.

The most interesting features of ferroelectric materials arise from the way they respond to an external field. If we begin with a multidomain single crystal and apply an external field parallel to a crystallographic axis, those domains having their polarization in the direction of the field are unaffected. However, domains oriented perpendicular to or opposite to the field are

unstable. Within these, new domains parallel to the field will nucleate and grow until ultimately the single crystal contains only one domain oriented with the field. The polarization curve will exhibit a hysteresis as shown in Fig. 12.8. At point A on the *P-E* curve, only one domain would exist and would be aligned with the field. If E is decreased, the polarization will vary with E along the curve to point B. At this juncture there is still one domain but now it is opposed to the external field. If E is decreased further, new domains oriented with the field start to grow. At point C, the electric moment of the domains oriented with the field just cancels the moment of the domains oriented against the field. The field intensity corresponding to point C is called the *coercive force* and the polarization corresponding to A is called the *saturation polarization.*

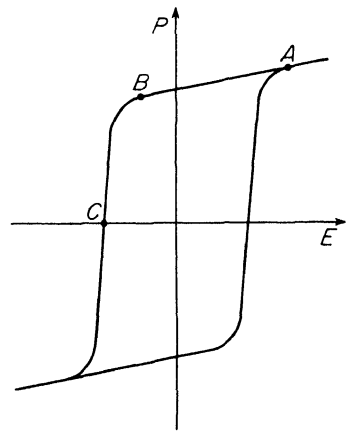


Figure 12.8 Hysteresis in $BaTiO_3$.

The shape of the hysteresis loop is greatly affected by lattice defects in a single crystal and by crystal size in a polycrystalline agglomerate. The slope

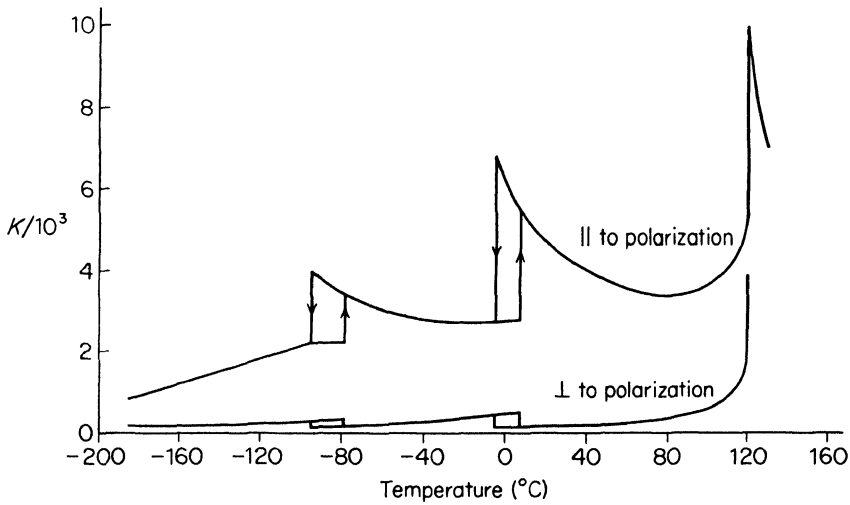


Figure 12.9 Variation of dielectric constant of $BaTiO_3$ with temperature. (After W. J. Merz, *Phys. Rev.* **76,** 1221 (1949).)

of the steep portion of the curve in Fig. 12.8 is governed by the ease with which new domains are nucleated and grow. This increases with increasing temperature up to the Curie temperature where the slope is almost infinite.

The phenomenology of this behavior was discussed previously in Chap. 6 (p. 129). Because of the hysteresis in the *P-E* curve, the dielectric susceptibility can have meaning only if no change in the domain structure of the sample is brought about by the measurement.

The dielectric constant and the dielectric susceptibility of $BaTiO_3$ are practically identical and are both very high. The variation in K with temperature is shown in Fig. 12.9.

Barium titanate is piezoelectric as well as ferroelectric. Among the devices in which it finds application are digital memory elements, nonlinear capacitors, dielectric amplifiers, and electromechanical transducers.

12.7 Antiferroelectricity

In many materials having the perovskite structure, the coupling through the oxygen octahedra cause adjacent lines of basic cells to be polarized in opposite directions. Notable among these is lead zirconate. In various temperature ranges it passes through two antiferroelectric, one ferroelectric, and one paraelectric phase. In strong fields, the antiferroelectric phases become ferroelectric, and substitution of some TiO_2 for ZnO_2 will also cause the antiferroelectric phases to become ferroelectric.

12.8 Piezoelectricity

If a material is asymmetric with respect to the distribution of positive and negative charges in its structure, it is to be expected that a strain would change this asymmetry and accordingly would change the internal polarization. Conversely, if the internal polarization were changed by an external electric field, it would be expected that dimensional changes in the crystal would result. This type of electromechanical interaction is called *piezoelectricity*.

The piezoelectric properties of a material can be characterized in terms of any one of the four rank three tensor quantities

$$\left.\frac{\partial D_i}{\partial \sigma_{jk}}\right)_{T,E=\text{const}} \qquad \left.\frac{\partial \epsilon_{ij}}{\partial E_k}\right)_{T,\sigma=\text{const}}$$
$$\left.\frac{\partial \sigma_{ij}}{\partial D_k}\right)_{S,\epsilon=\text{const}} \qquad \left.\frac{\partial E_i}{\partial \epsilon_{jk}}\right)_{S,D=\text{const}} \tag{12.16}$$

These are interrelated by the methods of Chap. 6.

Using $C_{ijk} = \partial E_i/\partial \epsilon_{jk})_{S,D}$ as an example, we can develop certain ideas which will be valid for all of these. A general rank three tensor will have 27 components. However, since $\epsilon_{jk} = \epsilon_{kj}$, only 18 will be distinct. These could be reduced to nine by a proper choice of coordinates, but if we adopt the convention of p. 147, there will be 18 components for the most symmetric case. The restrictions imposed by symmetry may cause some of these to be equal to each other or to zero. We will illustrate this by considering quartz.

Quartz has a trigonal trapezohedric structure which is characterized by a fourfold axis of symmetry perpendicular to a twofold axis of symmetry. When using the conventional coordinate system, the components of the **C** tensor must remain invariant under the transformations prescribed by the matrix operators

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

After these are performed and Neumann's principle applied, we find that all but six of the tensor components are zero. These six are related to one another by

$$C_{132} = -C_{312}, \qquad C_{123} = -C_{321}, \qquad C_{213} = -C_{231}, \qquad C_{213} = C_{123}$$

Consequently, only two of them are independent.

Once the elastic constants, the dielectric susceptibility, the piezoelectric constants, and the variation of all of these with temperature have been evaluated, the response of a crystal of a prescribed size and shape to a specified electrical and mechanical environment can be computed. When a transducer is cut from the raw crystal it is possible to decide beforehand how the cuts should be oriented with respect to the crystallographic axes to maximize specified modes of electromechanical coupling, or to minimize the effect of temperature upon the behavior of the transducer.

Piezoelectric materials are used in a wide variety of applications involving electromechanical energy conversion. Crystals shaped to have a prescribed mechanical resonance frequency can be used as narrow-band electrical filters that pass electrical signals whose frequency coincides with the mechanical vibration frequency and reject all others. Conversely, crystals that are subjected to dynamic strains produce an electrical output, and are used in this manner in vibration instrument systems. Alternatively, crystals can be driven electrically to produce acoustical waves as in sonar systems. In each of these general areas the design and analysis of the piezoelectric transducer for a specific application is remarkable for the degree to which art can be eliminated from, and science applied to, the problem.

Applications. Obviously, a detailed treatment of even a few of the possible applications of piezoelectric devices is out of the question here, but some of the general aspects of any piezoelectric transducer problem can be delineated fairly readily.

Let us suppose that the thermodynamic state of a piezoelectric transducer is determined either by the extensive parameters, temperature, stress and field intensity, or by the intensive parameters, entropy per unit volume, strain and polarization per unit volume. Then the various parameters are inter-

related by either of the sets

$$dT = \frac{\partial T}{\partial S}\Big)_{P,\epsilon} dS + \frac{\partial T}{\partial P_i}\Big)_{S,\epsilon} dP_i + \frac{\partial T}{\partial \epsilon_{ij}}\Big)_{P,S} d\epsilon_{ij}$$

$$dE_i = \frac{\partial E_i}{\partial S}\Big)_{P,\epsilon} dS + \frac{\partial E_i}{\partial P_j}\Big)_{S,\epsilon} dP_j + \frac{\partial E_i}{\partial \epsilon_{ij}}\Big)_{P,S} d\epsilon_{ij}$$

$$d\sigma_{ij} = \frac{\partial \sigma_{ij}}{\partial S}\Big)_{P,\epsilon} dS + \frac{\partial \sigma_{ij}}{\partial P_k}\Big)_{S,\epsilon} dP_k + \frac{\partial \sigma_{ij}}{\partial \epsilon_{kl}}\Big)_{P,S} d\epsilon_{kl}$$

or

$$dS = \frac{\partial S}{\partial T}\Big)_{\sigma,E} dT + \frac{\partial S}{\partial E_i}\Big)_{\sigma,T} dE_i + \frac{\partial S}{\partial \sigma_{ij}}\Big)_{E,T} d\sigma_{ij}$$

$$dP_i = \frac{\partial P_i}{\partial T}\Big)_{\sigma,E} dT + \frac{\partial P_i}{\partial E_j}\Big)_{\sigma,T} dE_j + \frac{\partial P_i}{\partial \sigma_{ij}}\Big)_{E,T} d\sigma_{jk}$$

$$d\epsilon_{ij} = \frac{\partial \epsilon_{ij}}{\partial T}\Big)_{\sigma,E} dT + \frac{\partial \epsilon_{ij}}{\partial E_k}\Big)_{\sigma,T} dE_k + \frac{\partial \epsilon_{ij}}{\partial \sigma_{kl}}\Big)_{E,T} d\sigma_{kl}$$

The partial derivatives are the compliances and rigidities discussed in Chap. 6. In applying these to the analysis of a piezoelectric device, the conditions of operation must be prescribed. For the purpose of analysis it is particularly convenient if we can specify that the crystal is clamped ($\epsilon = 0$) or free ($\mathfrak{d} = 0$), operating isothermally ($dT = 0$) or isentropically ($dS = 0$), and is short-circuited ($\mathbf{E} = 0$) or open-circuited ($d\mathbf{P} = 0$). For example, let us compute the output voltage of a crystal subjected to a uniaxial stress. We specify that the stress varies with time sufficiently slowly that the inertial forces in the crystal are negligible and that the process can be considered isothermal. The output of the crystal feeds into an infinite input impedance.

We begin by choosing a coordinate system with the electrodes attached to faces perpendicular to the x_1 direction and specify that all stress components except σ_{32} are equal to zero. The compliances

$$\chi_{ij} = \frac{\partial P_i}{\partial E_j}\Big)_{\sigma,T} \qquad \text{and} \qquad d_{ijk} = \frac{\partial P_i}{\partial \sigma_{jk}}\Big)_{E,T}$$

are known for the conventional coordinate system of p. 147 and if we specify the orientation of the crystallographic axes, the components of $\boldsymbol{\chi}$ and $\mathbf{d}$ can be transformed to the coordinate system for the problem. Finally the problem conditions state that

$$d\mathbf{P} = 0 = \boldsymbol{\chi} \cdot d\mathbf{E} + \mathbf{d} \cdot d\mathfrak{d}$$

If we use the Kronecker delta and define a rank two tensor $\boldsymbol{\lambda}$ such that

$$\boldsymbol{\lambda} \cdot \boldsymbol{\chi} = \boldsymbol{\delta}$$

then

$$\boldsymbol{\delta} \cdot d\mathbf{E} = \boldsymbol{\lambda} \cdot (\mathbf{d} \cdot d\boldsymbol{\delta})$$

This expands into

$$E_1 = (\lambda_{11}d_{122} + \lambda_{12}d_{222} + \lambda_{13}d_{322})\sigma_{22}$$

from which the terminal voltage is

$$V = E_1 l$$

where l is the thickness in the x_1 direction.

12.9 Diamagnetism and paramagnetism

There is a similarity between the parameters used to define the electrical properties of a dielectric and those used to define the magnetic properties of para- and diamagnetic solids. In free space, the magnetic field intensity may be characterized by the torque that it would exert upon a unit magnetic dipole. Such a quantity would be a vector and is usually represented by **H**. When a solid is placed in a magnetic field, two processes may occur which either augment or diminish the field intensity inside the solid. Any unpaired electrons or nucleons that are present in the solid will have magnetic moments due to their spin. These tend to align with the field and to augment it. This tendency is opposed by the thermal energy in the solid, and the alignment is temperature dependent. In addition to this, the change in magnetic flux occurring when the solid is placed in the field will induce an electromotive force causing all those electrons unable to interact with the lattice to circulate in such a direction that they set up an opposing magnetic field. The net result of the operation of both factors is to cause the local magnetic field inside the solid to differ from the applied field. The incremental field due to the interaction of the applied field and the solid can be characterized as a vector quantity representing the magnetic dipole moment per unit volume of the solid. This is given the symbol **M**. The interaction between the solid and the applied field is a rank two tensor property of solid called the *magnetic susceptibility*. This is defined by the equation

$$\chi_m = \frac{\partial \mathbf{M}}{\partial \mathbf{H}}\bigg)_{\sigma,T} \tag{12.17}$$

When the electrons in a material interact with a magnetic field to augment its intensity inside the solid, the material is said to be *paramagnetic*. When they act to diminish the field in a solid, the material is said to be *diamagnetic*. Actually both processes act in a paramagnetic material, but the diamagnetic response is always much weaker than the paramagnetic response. It is often possible to separate the two by appropriate experimental techniques, and a

diamagnetic and a paramagnetic susceptibility can be determined independently of one another in a paramagnetic material.

The total torque acting upon a dipole inside the solid is

$$\mathbf{B} = \boldsymbol{\mu}_0 \cdot \mathbf{H} + \boldsymbol{\mu}_0 \cdot \mathbf{M} = \boldsymbol{\mu} \cdot \mathbf{H} \tag{12.18}$$

where $\boldsymbol{\mu}_0$ is the *permeability of free space* and $\boldsymbol{\mu}$ is the *permeability tensor* of the material.

According to the Langevin theory,[1] the effect of temperature upon the magnetic susceptibility is

$$\chi_m = \frac{C}{T} \tag{12.19}$$

where C is the Curie temperature.

12.10 Ferro- and antiferromagnetism

A number of solid materials undergo the magnetic equivalent of a polarization catastrophe. The magnetic characteristics of these materials are analogous to the dielectric characteristics of materials that undergo a spontaneous electrical polarization. Thus we find a domain structure, hysteresis in the B-H curves, and one or more Curie temperatures marking transitions between structures having different kinds of spontaneous polarization or no polarization at all.

By analogy with their dielectric equivalents, *ferromagnetic crystals* have domains that are spontaneously polarized in one direction. *Antiferromagnetic materials* are those having domains in which the spontaneous polarization forms antiparallel pairs of polarized basic cells.

With few exceptions, the list of ferro- and antiferromagnetic materials is confined to either the elements having atomic numbers between 23 and 29 or compounds of these (See Table 12.2). These are all characterized by the fact that the $3d$ and $4s$ electronic states in the isolated atoms are very close in energy and the bands due to these electronic states overlap in the elemental solids. Following this clue, Mott and Jones postulated a quantum mechanical coupling between the spins of these electrons. Oddly enough this coupling is electrostatic rather than magnetic and leads to an exchange energy similar to the resonance effect described in Chap. 5. When the exchange energy is negative, this coupling causes a spontaneous alignment of electron spin. The exchange energy is negative only in a certain range of lattice parameters, and only a portion of the elements with some electrons in the $3d$ and $4s$ states also have their lattice parameters within the range that leads to coupling between

[1] Charles Kittel, *Introduction to Solid State Physics* (2nd ed., New York: John Wiley & Sons, Inc., 1953), p. 213.

electron spins. When these dimensions are achieved by alloying, combinations of elements in an alloy may be ferromagnetic even though the individual constituents of the alloy are not. The Heussler alloys illustrate this point rather well. None of the constituents is ferromagnetic by itself, but when alloyed, the dimensional requirements for strong parallel spin coupling are met and the alloy itself is ferromagnetic. This same result is accomplished in certain compounds; e.g. elemental chromium is not ferromagnetic but CrO is.

Table 12.2 *Ferro- and antiferromagnetic materials*

Material	*Curie temperature* (°*K*)
Ferromagnetic elements	
Fe	1043
Co	1400
Ni	631
Gd	289
Dy	105
Ferromagnetic alloys	
Cu_2MnAl	603
Cu_2MnIn	506
CrTe	336
Ferromagnetic compounds	
$MnOFe_2O_3$	783
$FeOFe_2O_3$	848
$MgOFe_2O_3$	583
Antiferromagnetic elements	
Mn	100*
Cr	475
Antiferromagnetic compounds	
MnO	122
$FeCl_2$	23.5

* Approximately.

Because of the influence of the lattice dimensions upon the spin coupling, the onset of spontaneous magnetic polarization will follow certain preferred crystallographic directions. For example, in iron and nickel, both of which are cubic, the preferred directions are ⟨100⟩, ⟨010⟩, and ⟨001⟩ for iron and ⟨111⟩, ⟨11$\bar{1}$⟩, and ⟨1$\bar{1}$1⟩ for nickel. We might expect that a multiplicity of directions of spontaneous polarization would lead to domain structures in ferromagnetic materials similar to those observed in ferroelectric materials. This is the case, but there are some important differences between ferroelectric and ferromagnetic domains. Perhaps contrasting them will lead to a better understanding of both. One of the major differences is that spontaneous electric polarization is necessarily accompanied by a structural change in the basic cell, whereas no such structural change is a necessary adjunct to spontaneous magnetic polarization. As a consequence, relatively high strain fields exist around the domain wall in a ferroelectric material. The wall

marks a sharp transition between regions having different polarization directions and the walls are difficult to move through the crystal.

In contrast, ferromagnetic polarization may produce a relatively small contraction of the basic cell along the direction of polarization, but this is a consequence rather than a cause of polarization. A ferromagnetic domain wall has only a relatively weak strain field associated with it and the wall will move rather easily. A further difference is that a ferromagnetic domain wall is several hundred Ångstrom units thick and the direction of polarization inside the wall changes continuously from that of the domain on one side to that of the domain on the other side.

The second major difference is best illustrated by considering two specific examples: $BaTiO_3$, cubic in its paraelectric state; and iron, cubic in both the para- and ferromagnetic states. Barium titanate can be polarized only in the

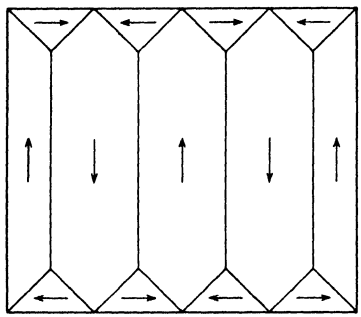


Figure 12.10 Schematic of magnetic domain structure in which the domains are arranged to minimize the energy stored in the magnetic field.

$\langle 100 \rangle$, $\langle 010 \rangle$, or $\langle 001 \rangle$ direction. Iron polarizes spontaneously in these directions, but an external field, even from a neighboring domain, can cause it to polarize in the $\langle 110 \rangle$, $\langle 011 \rangle$, $\langle 101 \rangle$, $\langle 11\bar{1} \rangle$, and $\langle 1\bar{1}1 \rangle$ directions. Indeed a ferromagnetic domain can be polarized in any arbitrary direction if the external field is sufficiently strong.

The relative mobility of domain walls in ferromagnetic materials, and the fact that they can also be polarized in secondary directions, makes it possible for the domains in a crystal to rearrange themselves in such a manner that the energy stored in their magnetic fields is a minimum. For example, the domain configuration shown in Fig. 12.10 provides a low-reluctance path confining the magnetic field almost entirely within the crystal, and the energy stored in the magnetic field is considerably smaller than it would have been had the crystal been composed of a single domain.

In the presence of an external magnetic field, new domains are not nucleated. Instead, at low field intensities, the domain walls move in such a way that the favorably oriented domains grow at the expense of the less

favorably oriented domains. At higher fields, the polarization direction inside a domain can rotate until the direction of polarization is aligned with the field. This is shown schematically in Fig. 12.11.

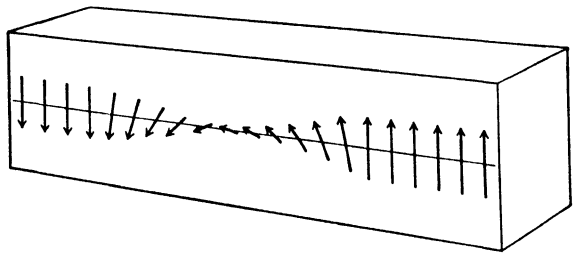


Figure 12.11 Domain wall structure in ferromagnetic materials.

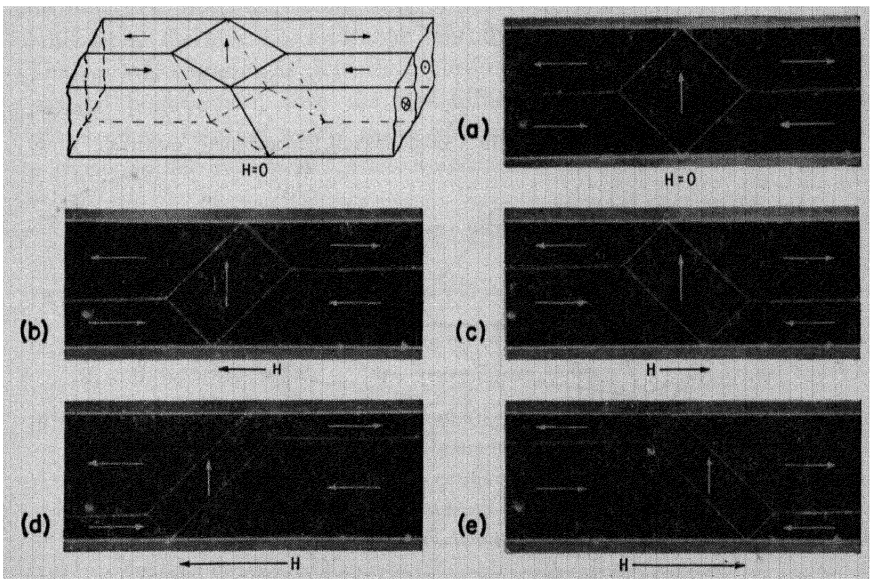


Figure 12.12 This sequence of photographs shows magnetic domains in a single crystal of iron. The external field intensity is zero in (a). The sequences (b), (d), and (c), (e) show the influence of increasing external field upon the growth of favorably oriented domains. (Courtesy of R. W. De Blois and C. K. Graham, Jr., General Electric Research Laboratory, Schenectady, N. Y.)

Figure 12.12 shows a set of five magnetic domains in a single crystal of iron. Two pairs are opposed and the fifth domain lies in a transverse direction. The sequence shows the effect of increasing external field in a direction perpendicular to the transverse domain. Note that the domain wall moves

in manner permitting the favorably oriented domains to grow at the expense of the transverse and anti-parallel ones.

The *B-H* curve of a ferromagnetic material exhibits a hysteresis which is characterized by a coercive force and a saturation polarization. Both of these vary with the direction of the field with respect to the crystallographic axes.

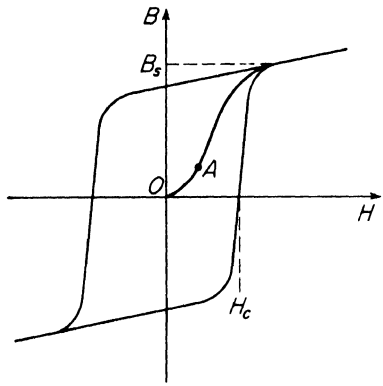


Figure 12.13 Hysteresis loop in a single crystal ferromagnetic material.

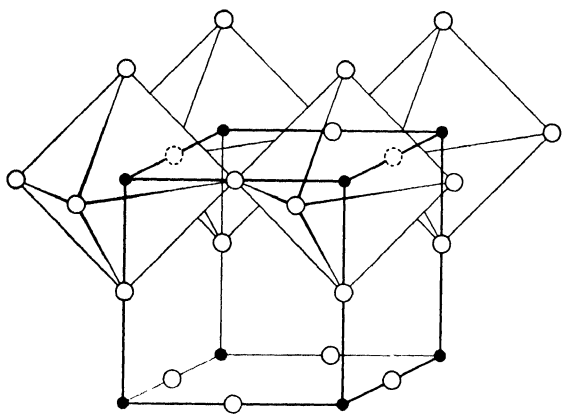


Figure 12.14 Spinel structure.

The behavior of the initial part of the *B-H* curve of Fig. 12.13 is interesting in that it is reversible along the *OA*. Up to field intensities corresponding to point *A*, the domain walls interact reversibly with crystal defects and will store increments of magnetic energy reversibly. At field intensities above that indicated by point *A*, the domain walls move through any impeding strain field, and if the field is alternated, the usual hysteresis loop results.

Any discussion of magnetic materials in this decade would be seriously incomplete if some attention were not given to the so-called *ferrites.* These are the magnetic spinel structures indicated in Table 12.2. A partial spinel structure is shown in Fig. 12.14. Its salient feature is that it consists of cubic close-packed oxygen ions with the smaller di- and tri-valent metallic ions in the interstices. There are two types of interstitial sites present—one surrounded tetrahedrally by oxygen ions, and one in which there are six nearest neighboring oxygen ions. Let us label these sites A and B respectively. There are 64 A and 32 B sites in a basic cell. In each basic cell there are one divalent ion and two trivalent ions for each four oxygen ions. Thus in a basic cell only 24 of the interstitial sites will be occupied, 16 of which are B sites. In most ferrites the divalent ions occupy the B sites and the trivalent ions are distributed equally between A and B sites. Strong coupling occurs only between A and B sites and no coupling will occur if the divalent ion has no unpaired electrons. Even when ions having unpaired electrons occupy both sites, there is no assurance that the crystal will be ferromagnetic. The normal situation is that some portion of the coupling results in antiparallel polarization, with the remaining coupling resulting in parallel polarization. In other words, the crystal may be partially ferromagnetic and partially antiferromagnetic. These materials are given the name *ferrimagnetic.*

PROBLEMS

12.1 Discuss the difference between **P** and **D** and between **K** and χ_e.

12.2 Outline the experimental procedure required to evaluate the components of **K** in a dielectric material.

12.3 Crystal classes 28 and 31 are in the cubic system and are, consequently, isotropic with respect to **K** and χ_e. The members of this class, however, could be piezoelectric. Even though they lack a center of symmetry, they have other symmetry elements which prevent them from being pyroelectric.

(a) If no magnetic field is present, which parameters must be considered in defining the dielectric constant?

(b) Distinguish between the clamped dielectric constant and the unclamped dielectric constant. Which of these is a true compliance and how are the two related to one another?

12.4 Compute the strain induced in a crystal by imposing the electric field $E_1 = 10^6$ volts/cm., $E_2 = 5 \times 10^5$ volts/cm., and $E_3 = 0$ and the stress $\sigma_{11} = 10^7$ dynes/cm., $\sigma_{22} = \sigma_{33} = 0$, $\sigma_{12} = 5 \times 10^9$ dynes/cm., and $\sigma_{23} = \sigma_{13} = 0$. The components of the piezoelectric compliance and the elastic compliance tensors are given below:

$$T_{123} = T_{213} = T_{312} = T_{132} = T_{231} = T_{321} = 4 \times 10^{-3} \text{ cm. volt}^{-1}$$

All other components of the piezoelectric compliance = 0.

$$S_{i(iii)} = 5 \times 10^{-12} \text{ cm}^2/\text{dyne} \qquad i = 1,2,3$$

$$S_{i(i)j(j)} = 7 \times 10^{-12} \text{ cm}^2/\text{dyne} \qquad i, j = 1,2,3$$

$$S_{i\,j(ij)} = 6 \times 10^{-12} \text{ cm}^2/\text{dyne} \qquad i, j = 1,2,3$$

All other components of the elastic compliance = 0.

12.5 Classify $BaTiO_3$ according to its symmetry when it is in the ferroelectric state that is stable at room temperature.

APPENDIX
A

Independent Components of Rank Three Tensor Properties where $T_{ijk} = T_{ikj}$

Class 1:

$$\begin{bmatrix} T_{111} & T_{122} & T_{133} & T_{123} & T_{113} & T_{112} \\ T_{211} & T_{222} & T_{233} & T_{223} & T_{213} & T_{212} \\ T_{311} & T_{322} & T_{333} & T_{323} & T_{313} & T_{312} \end{bmatrix} \qquad 1$$

Class 2:

Center of symmetry $\therefore\ T_{ijk} = 0 \qquad \bar{1}$

Class 3 (plane of symmetry perpendicular to x_3 axis):

$$\begin{bmatrix} T_{111} & T_{122} & T_{133} & 0 & 0 & T_{112} \\ T_{211} & T_{222} & T_{233} & 0 & 0 & T_{212} \\ 0 & 0 & 0 & T_{323} & T_{313} & 0 \end{bmatrix} \qquad m$$

Class 4 (x_3, is a twofold axis):

$$\begin{bmatrix} 0 & 0 & 0 & T_{123} & T_{113} & 0 \\ 0 & 0 & 0 & T_{223} & T_{213} & 0 \\ T_{311} & T_{322} & T_{333} & 0 & 0 & T_{312} \end{bmatrix} \qquad 2$$

Class 5:

Center of symmetry $\therefore\ T_{ijk} = 0 \qquad 2/m$

Class 6 (reflection planes perpendicular to x_1 and x_2 axes):

$$\begin{bmatrix} 0 & 0 & 0 & 0 & T_{113} & 0 \\ 0 & 0 & 0 & T_{223} & 0 & 0 \\ T_{311} & T_{322} & T_{333} & 0 & 0 & 0 \end{bmatrix} \qquad mm$$

Class 7 (all axes twofold axes of symmetry):

$$\begin{bmatrix} 0 & 0 & 0 & T_{123} & 0 & 0 \\ 0 & 0 & 0 & 0 & T_{213} & 0 \\ 0 & 0 & 0 & 0 & 0 & T_{312} \end{bmatrix}$$

222

Class 8 (reflection planes perpendicular to all three crystallographic axes):

$$\begin{bmatrix} 0 & 0 & 0 & T_{123} & T_{113} & 0 \\ 0 & 0 & 0 & -T_{113} & T_{123} & 0 \\ T_{311} & -T_{311} & 0 & 0 & 0 & T_{312} \end{bmatrix}$$

mmm

Class 9:

$-T_{ijk} = 0$ because of rotation reflection axis

$\bar{4}$

Class 10 (x_3 is a fourfold axis):

$$\begin{bmatrix} 0 & 0 & 0 & T_{123} & T_{113} & 0 \\ 0 & 0 & 0 & T_{113} & -T_{123} & 0 \\ T_{311} & T_{311} & T_{333} & 0 & 0 & 0 \end{bmatrix}$$

4

Class 11:

Center of symmetry $\therefore T_{ijk} = 0$

4/m

Class 12 (x_3 is a fourfold axis; x_1, x_2 are twofold axes):

$$\begin{bmatrix} 0 & 0 & 0 & T_{123} & 0 & 0 \\ 0 & 0 & 0 & 0 & T_{123} & 0 \\ 0 & 0 & 0 & 0 & 0 & T_{312} \end{bmatrix}$$

42m.

Class 13 (x_3 is a fourfold axis and is parallel to intersection of two orthogonal reflection planes):

$$\begin{bmatrix} 0 & 0 & 0 & 0 & T_{113} & 0 \\ 0 & 0 & 0 & T_{113} & 0 & 0 \\ T_{311} & T_{311} & T_{333} & 0 & 0 & 0 \end{bmatrix}$$

4mm

Class 14 (x_3 is a fourfold axis; x_1, x_2 are twofold axes):

$$\begin{bmatrix} 0 & 0 & 0 & T_{123} & 0 & 0 \\ 0 & 0 & 0 & 0 & -T_{123} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

42

Class 15:

Center of symmetry, $T_{ijk} = 0$

4/mmm.

Class 16 (x_3 is a threefold axis):

$$\begin{bmatrix} T_{111} & -T_{111} & 0 & T_{123} & T_{113} & -T_{222} \\ -T_{222} & T_{222} & 0 & T_{113} & -T_{123} & -2T_{111} \\ T_{311} & T_{311} & T_{333} & 0 & 0 & 0 \end{bmatrix}$$

3

Class 17:

Center of symmetry, $T_{ijk} = 0$ $\bar{3}$

Class 18 (x_3 axis is threefold axis; plane of symmetry is parallel to x_2 axis)

$$\begin{bmatrix} 0 & 0 & 0 & 0 & T_{123} & -2T_{222} \\ -T_{222} & T_{222} & 0 & T_{123} & 0 & 0 \\ T_{311} & T_{311} & T_{333} & 0 & 0 & 0 \end{bmatrix}$$ m J

Class 19 (x_3 is threefold axis; x_2 is twofold axis):

$$\begin{bmatrix} T_{111} & -T_{111} & 0 & T_{123} & 0 & 0 \\ 0 & 0 & 0 & 0 & -T_{123} & -2T_{111} \\ T_{311} & T_{311} & T_{333} & 0 & 0 & 0 \end{bmatrix}$$ 32

Class 20:

Center of symmetry $\therefore T_{ijk} = 0$ $\bar{3}m$.

Class 21 (x_3 is sixfold axis):

$$\begin{bmatrix} 0 & 0 & 0 & T_{123} & T_{113} & 0 \\ 0 & 0 & 0 & T_{113} & -T_{123} & 0 \\ T_{311} & T_{311} & T_{333} & 0 & 0 & 0 \end{bmatrix}$$ 6

Class 22 (x_3 is threefold axis parallel to reflection plane):

$$\begin{bmatrix} T_{111} & -T_{111} & 0 & 0 & 0 & -2T_{222} \\ -T_{222} & T_{222} & 0 & 0 & 0 & -2T_{111} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$ $\bar{6}$

Class 23:

Center of symmetry $\therefore T_{ijk} = 0$ 6/m.

Class 24 (x_3 is sixfold axis; planes of symmetry parallel to x_3 and x_2 axes):

$$\begin{bmatrix} T_{111} & -T_{111} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -2T_{111} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$ $\bar{6}m2$ a

Class 25 (x_3 is sixfold axis; plane of symmetry parallel to x_2 axis):

$$\begin{bmatrix} 0 & 0 & 0 & 0 & T_{113} & 0 \\ 0 & 0 & 0 & T_{113} & 0 & 0 \\ T_{311} & T_{311} & T_{333} & 0 & 0 & 0 \end{bmatrix}$$ 6mm

Class 26 (x_3 is sixfold axis; x_1 is twofold axis):

$$\begin{bmatrix} 0 & 0 & 0 & T_{123} & 0 & 0 \\ 0 & 0 & 0 & 0 & T_{123} & 0 \\ 0 & 0 & 0 & 0 & 0 & T_{123} \end{bmatrix}$$ 62

Class 27:

Center of symmetry $\therefore T_{ijk} = 0$ 6/mmm

Classes 28 and 31 (all axes are twofold):

$$\begin{bmatrix} 0 & 0 & 0 & T_{123} & 0 & 0 \\ 0 & 0 & 0 & 0 & T_{123} & 0 \\ 0 & 0 & 0 & 0 & 0 & T_{123} \end{bmatrix}$$

23 ; 43.

Classes 29, 30, and 32:

Centers of symmetry $\therefore T_{ijk} = 0$

m3 ; 43m ;

APPENDIX C

Periodic Table of the Elements

Periodic Table of the Elements

1 A							VIII A
H 1 1.0080	II A	III A	IV A	V A	VI A	VII A	He 2 4.003
Li 3 6.940	Be 4 9.013	B 5 10.82	C 5 12.01	N 7 14.008	O 8 16.000	F 9 19.00	Ne 10 20.183
Na 11 22.997	Mg 12 24.32	Al 13 26.98	Si 14 28.09	P 15 30.975	S 16 32.07	Cl 17 35.46	A 18 39.94
K 19 39.100	Ca 20 40.08	Ga 31 69.72	Ge 32 72.60	As 33 74.91	Se 34 78.96	Br 35 79.916	Kr 36 83.8
Rb 37 85.48	Sr 38 87.63	In 49 114.76	Sn 50 118.70	Sb 51 121.76	Te 52 127.61	I 53 126.91	Xe 54 131.3
Cs 55 132.91	Ba 56 137.36	Tl 81 204.39	Pb 82 207.21	Bi 83 209.00	Po 84 (210)	At 85 (210)	Rn 86 222
Fr 87 223	Ra 88 226.05						

I B	II B	III B	IV B	V B	VI B	VII B
		Sc 21 44.96	Ti 22 47.90	V 23 50.95	Cr 24 52.01	Mn 25 54.93
Cu 29 63.54	Zn 30 65.38	Y 39 88.92	Zr 40 91.22	Ni 41 92.91	Mo 42 95.95	Tc 43 99.1
Ag 47 107.88	Cd 48 112.41	57–71 *	Hg 72 178.6	Ta 72 108.88	W 74 183.92	Re 75 186.31
Au 79 197.2	H 80 200.61	89– 101 **				

VIII B		
Fe 26 55.85	Co 27 58.94	Ni 28 58.69
Ro 44 101.7	Rh 45 102.91	Pd 46 106.7
Os 76 190.2	Ir 77 193.1	Pt 78 195.23

* Lanthanide Series

La 57 138.92	Ce 58 140.13	Pr 59 140.92	Nd 60 144.27	Pm 61 (147)	Sm 62 150.43	Eu 63 152.0	Gd 64 156.9	Tb 65 159.2	Dr 66 162.46	Ho 67 164.94
Er 68 167.2	Tm 69 169.4	Yb 70 173.04	Lu 71 174.98							

** Actinide Series

Ac 89 (227)	Th 90 232.12	Pa 91 231	U 92 238.07	Np 93 237	Am 95 (243)	Cm 96 (245)	Bk 97 (249)	Cf 98 (249)	E 99 (253)	Fm 100 (254)
Mu 101 (256)										

INDEX

ACCEPTORS, 221
Activated processes, 201–203
Activation energy, 201–204
Alloys, 69, 182
  Heussler, 265
  Hume-Rothery rules, 109
Amorphous solids, 54, 74–78, 199
Antiferroelectricity, 260
Antiferromagnetism, 264
Areal density of atoms on crystallographic planes, 58–59, 174, 177
Atomic spectra, 81, 84–94
Atomic structure, 84–94

BAND, conduction, 98
Band theory, 97–100, 208–209
  correlation with zone theory, 109–111
  covalent solids, 98–99, 212–13
  metals, 99–100, 215–17
  semiconductors, 98–99, 217–22
Band, valence, 99
Basic cell, 55
Bohr's model, 81
Boltzman constant, 190, 200, 203, 212
Bond energy, 95
Bonding, 80–113
  covalent crystals, 94–99
  ionic, 111–13
  ionic-covalent, 111
  metallic crystals, 97–100
Bonding (*cont.*):
  van der Waals, 111
Bonds, hybrid, 95–97
Bragg reflection, 208
Brillouin planes, 105, 109
Brillouin surface, 106, 108, 189, 191, 207
Brillouin zone, 104–110, 189, 191, 207, 215–16
  volume of, 107–109, 193
Bubble raft, 68
Bulk modulus, adiabatic, 121
  isentropic, 119

CENTER of symmetry, 137, 138, 140–42, 149, 271–73
Characteristic equation, 34
Characteristic temperature, (*see* DeBye)
Characteristic values, 34
Charge carriers, concentration of, 213, 216–17, 242
  effective mass of, 206–210, 216
  mobility of, 214–15, 219, 221, 241
    table, 220
  in semiconductors, 218–24
  thermal conduction, 199–201
Coefficients, (*see* specific type)
Coercive force, 259, 268
Co-factor of determinant, 28
Compliance, 116
  mathematical nature, 118–28

Compliance (*cont.*):
matrix, 118
Concentration gradient, 13, 222, 237
Conductivity, (*see* specific type)
Contact potential, 235
Continuity equation, 21
Coordinates, Cartesian, 4–6, 56, 102, 104
principal, 44
spherical, 84–85
transformation of, 29–31, 40, 44
Coordination number, 62–63
Covalent bond, 94–99, 220
Creep, 172, 184–85
Critical shear-stress, (*see* Shear-stress, critical)
Crystal classes, 136–39
table, 137
Crystal defects, 63–69
Crystal growth, 65, 67, 68, 180
Crystallographic axes, 56–57, 70–72, 74, 178, 182–83
relation to principal axes, 147–48
table, 61–62
Crystallographic planes, 57–59, 174–76
Crystal momentum, 207–208, 214–16
Crystal structure, 55
Crystal symmetry 136–40
Crystal systems, 59–62
table, 61–62
Crystal twins, 71–72
Curie temperature, 127, 255–59, 264
table, 257
Current densities, (*see* Electrical, Entropy, Mass)

Davisson-Germer experiment, 83
De Broglie equation, 83, 102, 207
DeBye temperature, 193, 199
table, 193
De Hass-van Alphen effect, 245–46
Density of states function, 100, 106, 206–209, 212, 217–18, 234
Determinant, 28, 34, 44
Diamagnetism, 263
Diamond lattice, 98, 138
Dielectric constant, 111, 118, 251–54
Dielectric displacement, 116, 122, 250–51
Dielectric permittivity, 111, 118, 142, 251–54
Dielectric susceptibility, 251, 260, 262
Diffusivity, 13, 69
Diode, solid state, 99, 222–23
Dislocations, edge, 64, 66, 180
motion of, 66–67, 171, 180–82
semi-, 64, 65
spiral, 65, 66, 67, 180
Dislocation theory, 63–69, 179–85
charge carrier mobility, 214–15
and slip, 180–83
Distortion dyad, 156, 167–70
Divergence, 16
Divergence theorem, 17, 167, 196
Domains, ferroelectric, 256–59
and ferromagnetic, 265–68
Donors, 220–21
Dyadic product, 47, 123

Effective mass, 208–210, 214–19
Eigenvalues, 34
Einstein theory of specific heat, 190
Elastic compliance, 118, 158–61, 162, 163, 166
Elastic constants, 161–63, 165, 166, 169–70
semi-matrices, 275–77
tables, 162, 171
Elasticity, 155–72
Elastic limit, 172–73, 180–82
Elastic moduli, technical, 163–66, 169–70, 183–84, 193
Elastic rigidity, 118–19, 158–62, 166–69
table, 162
Electrical conductivity, 69, 142, 213, 225–28
in metals, 215–16
mobility of charge carriers, 214–15, 241
in semiconductors, 217–22
and thermal conductivity, 199–207
Electrical current density, 213, 225–26
Electrical resistivity, 214, 225–29, 240–41
Electric field intensity, 116, 122, 129, 225–29, 240–41, 249–50
Electro-caloric effect, 123
Electron density distribution function, (*see* Orbital)
Electron diffraction, 83
by lattice, 104–107, 116
Electronic configuration of elements, 90–93
Electron mobility, 214–15
Electrons, concentration of, 200, 213, 218, 221
Electron wave-particle dualism, 83

Electron waves, 83, 102–107, 189, 206–208
  spectral distribution, 103, 207–208
  velocity of propagation, 207–208
Electrostriction, 125
Energy gap, 99, 100, 110, 217–21
  table, 220
Energy, internal, of solids, 158–59, 189–92
Energy levels, 85
  covalent solids, 98
  hydrogen, 86
  metals, 215–17
  notation, 88
  semiconductors, 217–18
Enthalpy gradient, 197
Entropy, as parameter, 116, 121–22, 126–27
Entropy current density, 18–19, 128–29, 225
Equilibrium properties, 116–28
Equipartition of energy theory, 190
Ettinghausen effect, 245
Exchange energy, 95–96
Exclusion principle, 88
Extensive parameters, 116–28

FERMI-Dirac distribution function, 210–13 218, 234
Fermi energy or level, 212–13, 215–16, 218, 220, 234–37
Fermi surfaces, 209, 210, 246
Ferrimagnetism, 269
Ferrites, 269
Ferroelectricity, 257–60
Ferromagnetism, 129–31, 264–69
Function, (*see* specific type)

GALVANOMAGNETIC properties, 243
Gauss' theorem, 20–21
Glasses, 75, 205
Glide lines, 173
Gradient, 12–15 (*see* specific type)
  steady-state system, 128–29, 225–29
Grain boundaries, 69–74, 181–83, 188

HALL amplifier, 244–45
Hall coefficient, 239
  in magneto-resistance, 242–45
Hall effect, 239–45
Hall field, 239–40, 245
Holes, 209–210, 216–18, 236
Hume-Rothery rules, 109
Hybridization of bonds, 95–97
Hydrogen atom, quantum states, 84–88
Hysteresis, 129–31, 259, 268

IONIC bonding, 111–13
Ionic-covalent bonding, 111
Integrals, 16–17
Intensive parameters, 116

KELVIN relationships (*see* Thompson)
Kronecker delta, 5, 48

LATTICE, *bcc*, 55, 60, 62
  Brillouin zones, 105–106
Lattice, definition, 55
Lattice defects, 63–69, (*see* Dislocations)
Lattice diffraction, 104–107, 110, 207
Lattice, *fcc*, 55, 60, 62
  Brillouin zones, 105–106, 108
  slip systems, 175–78
Lattice, *hcp*, 55, 60, 62
  Brillouin zones, 107–108
Lattice, thermal energy, 191–92
Lattice vibrations, 187–89
Leduc-Righi effect, 245
Levi-Civita density, 8, 16, 49
Linear density of atoms in crystallographic directions, 58–59, 174, 177
Line integrals, 15
Lorentz force, 239
Lorentz number, 200–201

MADELUNG constants, 112–13
Magnetic field intensity, 116, 126–27, 129–31, 263–64
  and domain wall motion, 266–68
  and hysteresis in ferromagnetism, 268
Magnetic induction, 116, 263
Magnetic permeability compliance, 118, 129–31, 142
Magnetic susceptibility, 263–64
Magnetocaloric compliance, 124
Magnetoresistance, 242–45
Magnetostriction coefficient, 120
Many-valley solids, 209–210
Mass current density, 129
Material properties, classification, 115
  Class IV, 132–33
  Class I, equilibrium, 116–28

Material properties (*cont.*):
Class III, 129–31
Class II, steady state, 128–29
Matrices, elastic constants, 275–77
rank three tensor components, 271–74
rank two tensor components, 148
as rotation operators, 29–31
transformation, 140
types, 25, 27–29, 32
Matrix, compliance, 118
diagonalization, 31–35
inverse, 27
notation, 24–25
operations, 25–26
rigidity, 119
transpose, 28
unit, 27
Metals, band theory, 99, 100, 215–17
bonding in, 99–100
electrical conductivity, 215–17
thermal conductivity, 199–201
zone theory, 108–109
Metastability, 203–205
Michelson-Morley experiments, 80
Miller indices, 57–58
Molecular solids, 97, 111
Momentum, angular, 87
crystal, 207–208, 214–16
Momentum space, 4, 102, 208–210, 215

NERST effect, 245
Neumann's principle, 141, 159, 195, 240, 251, 261

ORBITALS, 85–88, 95
anti-bonding, 95
hybrid, 97
p-, 87, 97
s-, 95, 97
Orthogonality, 5

PARAMAGNETISM, 263
Parameters, extensive and intensive, 116–28
Pauli exclusion principle, 88–89
Peltier coefficient, 230–32, 237–38
Permeability, magnetic, 142, 264
saturation, 129–31
zero point, 129–31
Permittivity, 142
dielectric, 250–52
Periodic table, 279
Phonon, concentration, 190
definition, 188
energy, 189–94
mobility, 197–99
scattering factors, 197–99
spectral distribution function, 189–93
velocity of propagation, 191–92, 197–99
waves, 188–89, 191–92, 197–99
Photoelectric effect, 80, 84
Photon, 84–85
Piezoelectricity, 72, 118, 125–26, 148–49, 260–63
matrices of tensor components, 271–74
Piezoelectric stress coefficient, 118, 260, 262
Piezomagnetic stress coefficient, 120
Planck, black body radiation, 190
Planck's constant, 81, 82
Plasticity, 172–84
Poisson's ratio, 164–66, 170
Polarization, 123–24
in dielectrics, 249–57
domains, 256–57
in ferroelectrics, 257–59
in ferromagnetism, 264–69
domains, 265–68
relaxation time, 252–54
Polycrystalline materials, 54, 69–74, 203–205
plasticity, 173–74, 181–85
Polymers, 75–78
Pressure, as parameter, 116
Principal axes, 44
of tensor properties, 147–48
Properties of materials (*see* Material properties)
Pyroelectricity, 123, 141, 257
Pyroelectric rigidity, 119
Pyromagnetism, 120, 124, 141

QUANTUM numbers, azimuthal, 86
electron spin, 87
magnetic, 87
principal, 85
Quantum states, degenerate, 87, 103, 108
distribution in energy, 84–87, 102–103, 106–107
notation, 88

RATE processes, 201–203
Relaxation time, 200, 252–54
Resistivities, (*see* specific type)
Resonance, 95–96
Richardson work function, 235–36
Rigidities, 119 (*see* specific type)
  mathematical nature, 121–28
Rigidity matrix, 119
Rotation dyad, **R**, 157
Rotation operators, 29–31
Rutherford's experiment, 81

SCALAR, fields, 11–12
  gradient, 12
  -tensor effects, 124
  differentiation with vector, 8–9
Schroedinger wave equation, 83
Second-order effects, 126–28
Seebeck coefficient, 230, 232–34
Segregation, 72–74
Semiconductors, band theory, 98–99, 217–22
  intrinsic, 218
  n-, 220–21
  p-, 221–22
  zone theory, 109–110
Shear modulus, 164–65, 170
Shear-stress, critical, in slip, 174–78
  perfect single-crystal crystals, 178–80
Slip, 172–81
Slip systems, 174–78
Sound waves, 167, 188
Space, (*see* Momentum)
  n-, 102, 189, 207
  phase, 4, 102
Specific heat, 118, 187, 190–94
Strain, 155–57
  dynamic, and speed of sound, 167–70
Strain field, 66–67, 69, 181–82
Strain hardening, 181–82
Strain, as system parameter, 116
  in thermal expansion, 124, 194–95
Strain tensor, 44, 157, 168
Strain wave (*see* Phonon), 167–69, 181
Stress, 31–32, 66–67, 155, 157, 166, 172, 184–85
  as system parameter, 116
  tensor, 38–39, 122
Structural scattering factor, 197–98
Summation convention, 1–2, 40
Surface integral, 15–16
Symmetry of crystals, 136–52, 159–61, 164–65, 271–74
Symmetry elements, 136–39
Symmetry, of matrices, 25
  of tensors, 43–45, 49–50
Symmetry transformations, 138
  matrices, 140

TECHNICAL elastic moduli, (*see* Elastic)
Temperature (*see* Curie and DeBye)
  gradient, 13–14, 129, 196–97, 225
  as parameter, 116
Tensor, anti-symmetric, 43
  notation, 40
  operations, 43, 45–48
  rank four properties, 158–61
  rank one properties, 123–24, 141–42
  rank three properties, 148–52
    matrices, 271–74
  rank two properties, 142–47
    matrix, 148
  symmetry, 43, 49
  transformations, 41–42
  -tensor effects, 125–26
  unit, 48–49
Tensors, principal axes, 44
Thermal conductivity, 13, 18–20, 69, 129, 142, 187, 195–201, 225–28
Thermal expansion coefficient, 118, 120, 124, 142, 158, 187, 194–95
Thermal properties, 187–205
Thermal resistivities, 129, 225–28
Thermal scattering factor, 197–98
Thermoelectric conductivity, 129, 225–28, 238
Thermoelectricity, 229–39
  mechanism of, 234–38
Thermoelectric resistivity, 225–28
Thermogalvanic effect, 245
Thermomagnetic effect, 245
Thompson coefficients, 231, 237–38
Thompson relationships, 231–32
Transformation matrices for symmetry elements, 140
Transformations, matrix, 29–31
  orthogonal, 5
  similarity, 32
  tensor, 40
Transient systems, 21, 195–96
Transistors, 223–24
Twinning, 71–72, 173–74

Uncertainty principle, 82
Unit matrix, 27
Unit tensor (*see* Kronecker delta, Levi-Civita density)
Unit vectors, 10

Van der Waals bonding, 97, 111
Valence band, 99
Valence and electronic structure, 96–97
Vector definition, 6
  divergence theorem, 17
  fields, 11–12
  operations, 47–48
  operators, 16–17
  properties, 123–24, 141–42
  as tensor, 40, 45
  -tensor effects, 124
  transformations, 5, 30, 32, 33
Vector wave number, 102
Velocity, light, 84
  sound, 167
Vibrations, lattice, 187–88
  modes, 187, 190
Volume integral, 16
Volume, as parameter, 116

Wave equation, 83
  solutions of, 101, 104
Wave function, 82
Wave-number vector, 102
Whiskers, 180
Work function, 235–36
Work hardening, 181–82

Yield point, 69, 71
Young's modulus, 164–66, 169, 170, 183

Zone, Brillouin (*see* Brillouin)
Zone theory, 97, 100–108, 206–210
  correlation with band theory, 109–111
  metals, 108–109